nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

EVOLVING THREAT

Applying evolutionary theory to the biology of tumour resistance and how to overcome it

PAGE 166

TECHNOLOGY

THE QUANTUM INTERNET

A to-do list for the ultimate communications network

PAGE 169

COMMERCIALIZATION

PROFIT AND LOSS ACCOUNT

Salutary lessons from CRISPR–Cas9 patent battle

PAGE 172

CANCER GENOMICS

MUTATION HOTSPOTS

Transcription initiation at odds with DNA repair

PAGES 181, 259 & 264

NATURE.COM/NATURE

14 April 2016 £10

Vol. 532, No. 7598

9 770028 083095 15

THIS WEEK

EDITORIALS

WORLD VIEW Did dislike of bland national foods drive obesity? **p.149**

DAM BUSTING Marathon journey of Amazon catfish under threat **p.150**

LAND-HO! SpaceX finally lands its rocket on a base at sea **p.152**

Breeding controls

Scientists must help to inform regulators wrestling with how to handle the next generation of genetically engineered crops.

On 6 April, activists gathered in Paris to protest against an emerging class of genetically altered crops. Regulators often classify these as the product of 'new breeding techniques' (NBTs) that are sometimes distinct from classical — and historically controversial — genetically modified (GM) varieties. But some protesters, such as those who joined the Friends of the Earth demonstration in Paris last week, are unconvinced by that argument. They call the new plants 'hidden GMOs'.

Around the world, regulators are struggling to decide how to adapt the existing rules for transgenic technology to plant varieties that have been engineered using cutting-edge methods (see page 158). Many have found that their classical regulatory triggers rely on definitions of 'transgenic' or 'genetically modified organism' (GMO) that no longer apply. And they are questioning whether some NBT crops need to be regulated at all.

It is a complex problem, and one that demands steady input from researchers who are familiar with the science behind the technology.

Both terms — NBTs and hidden GMOs — attempt to hold an umbrella over a wide range of methods. Some of them are neither new technologies nor breeding techniques; many do diverge significantly from classical GM technology. The terms often apply to crops engineered using enzymes called nucleases that can be targeted to alter a specific DNA sequence, creating mutations or inserting new sequences into the genome. The wildly popular CRISPR–Cas9 gene-editing technique, for example, falls into this class. But the term NBT also refers to methods for silencing genes using RNA interference, for creating mutations without using nucleases, and even for grafting a non-GM plant onto a GM rootstock.

Public and regulatory discussions sometimes lump these techniques together, but the plants they yield can differ widely. Some mutations that are edited into the genome already exist in wild plant relatives in nature. Should such crops be regulated as stringently as crops in which CRISPR–Cas9 has been used to insert a fresh sequence into the genome? What if the insertion were 2 DNA letters, or 200?

It is clearly a challenge to gather all of this under a coherent regulatory framework that does not over- or under-regulate NBT crops. There will be a push for simplification. Researchers should seize every opportunity to inform the process, and to ensure that the simplification does not distort oversight.

The approach to oversight of GM crops at the US Department of Agriculture shows how a regulatory system can stray from science. GM crop regulations at that agency depend on its authority to control plant pests and noxious weeds. It is a system that had some relevance to the first generation of such crops, many of which were designed using genetic elements from plant pathogens.

It is rapidly losing relevance in the face of NBTs. In more than two dozen cases, the agency has determined that a particular NBT plant variety does not fall under its purview for regulation because it does not entail the use of a plant pest and is unlikely to yield a noxious weed. These might have been scientifically sound decisions, but they were not made for scientifically sound reasons.

The agency is currently revisiting that regulatory structure. There is ample opportunity for scientists to participate: it has released a draft statement listing some of the regulatory possibilities, and the public can comment until 21 April (see go.nature.com/oftgcw). The US National Academies of Sciences, Engineering, and Medicine has convened a committee to evaluate future developments in biotechnology products, including engineered crops, and to examine how those developments could affect regulations. The report is likely to be influential, and scientists should take part in the discussions as much as possible.

"There is room for a healthy debate as to how these crops are regulated."

Such opportunities are not limited to the United States: participation in other regions may not be as direct but could still be influential. Rather than wait on a long-delayed report from the European Commission to guide regulators, the European Plant Science Organisation in Brussels, for example, has already issued statements and put together educational material regarding NBTs (see go.nature.com/vcedfo).

There is room for a healthy debate as to how these crops are regulated: some may advocate for more oversight, others may want to loosen the reins. But for that debate to be fruitful, it must be well informed. Scientists with an interest in this field have a duty to ensure that it is. ■

Under appeal

Don't get too excited about that successful appeal against a grant rejection.

Since last week, *Nature* has been running an informal poll on its website, with striking results. Almost half of the thousand or so scientists who responded did not realize that it can be possible to appeal when they have a grant application rejected.

The poll was prompted by the remarkable story of a UK lab that successfully challenged such a rejection, and was subsequently awarded a €5-million (US$5.7-million) grant. As we report on page 159, computer scientist Peter Coveney at University College London convinced the European Commission that it had made a mistake in turning down his bid to create a hub to apply computer models to biomedical data.

"If your research is in jeopardy as a part of poor decisions, then people should be prepared to challenge them," Coveney said, in a

rallying cry that will surely be applauded in labs across the world. What scientist does not feel wronged when their valuable contribution to society is not recognized and their application for funds spurned?

As inspiring as Coveney's victory may seem to the ranks of the downtrodden and unappreciated, his case is unlikely to produce a surge of similar appeal successes. For starters, the fact that many scientists who answered our online poll did not know about possible appeals processes has made little difference to them or to their fortunes. Many big funders, including several national agencies, don't allow appeals. Just like in sport, the referee's decision is final, however unjust it might seem. And for those agencies that do allow appeals (a good way to find out is to check the funder's website) any complaint must provide concrete evidence of an error. In Coveney's case, the European Commission had mistakenly marked his application down for including something it had asked for.

Oh, and don't call your appeal a complaint. As many agencies — even those that do permit appeals — make clear, they don't respond in the same way to complaints. (Some, however, do allow appeals against results of investigations into complaints.)

The US National Institute of Allergy and Infectious Diseases (NIAID), for one, says that it prefers to start by handling any appeal as a 'grievance', which is turned into an 'appeal' only if it cannot be resolved. In that case, an authorized organizational representative (AOR) must write to the relevant NIAID programme officer. If the programme officer, who maybe working with a scientific review officer, disagrees with the AOR, then NIAID will send the appeal to its advisory council, and then on to the National Institutes of Health's Center for Scientific Review. The principal investigator does not revise the rejected application, which is re-reviewed by either the same or a different scientific review group.

Confused? Any similarities between the complexity of some appeals processes and the way that rail companies, say, make the process to claim refunds for delayed services so complicated that most people don't bother are surely coincidental. And yet, back in 1987, an article in *The Scientist* pointed out that the formal process for appealing against rejected grants was "one of the best-kept secrets in the scientific community" and added, cryptically, that "science administrators seem content to leave it that way" (see go.nature.com/d99fc5).

"A successful appeal may not guarantee extra funds."

The secret is out now, thanks to Coveney's efforts, and the decision in his favour announced last month. He and his co-applicants hired a lawyer to help them to negotiate the appeals process, but then the European Commission is known for its tortuous bureaucracy. Some research funders do, at least at first glance, seem to make the appeals process more benign. The British Academy, for example, simply invites those rejected to write to the chief executive, and then, as a last resort, to the president. Science Foundation Ireland intriguingly allows spurned applicants to appeal on the grounds of a wide range of possible failings in its procedures including the "inappropriate consideration of rumour/hearsay" by grant reviewers.

Be warned, though: a successful appeal does not guarantee extra funds. In its policy on grant-application appeals, the Natural Sciences and Engineering Research Council of Canada (NSERC) states: "If NSERC concludes that a procedural error occurred during the review of the application, the resulting funding decision could be to leave the original decision unchanged, or to increase or decrease the level and/or duration of the grant or award." Have appeals ever looked so unappealing? ■

Destination Venus

Findings from the Akatsuki mission should rekindle interest in Earth's closest neighbour.

When the first robotic probe penetrated Venus's cloud-filled atmosphere in 1967, it was designed to float. At the time, the surface of Venus was a complete mystery, and the engineers behind the Soviet Venera 4 thought it might land in a vast ocean. Science-fiction writers had imagined tropical swamps, forests or water worlds beneath the clouds. Venus's mass, density and composition were all similar to Earth's, and it was our closest neighbour, so it looked like a good bet for native life and even human colonization.

Instead, Venera 4 was destroyed before it reached the surface. The readout from its descent, and from subsequent probes, revealed extreme pressure, searing temperatures close to 500 °C and an atmosphere that was 95% carbon dioxide. Even though Venus was originally very like Earth, perhaps even replete with oceans, a runaway greenhouse effect had turned it into a hellhole. No one, it seemed, would be going to holiday on Venus any time soon.

The discovery that the brightest body in the sky, bar the Sun and the Moon, is so hostile to life has helped to turn humanity's attention to Mars, our next-closest neighbour. Not only is the red planet a more viable candidate for an off-Earth base, it is much easier to study. On Venus, dense clouds of sulfuric acid mean that only radar can trace the surface from the air. Two rovers are trawling Mars right now, and more are in the pipeline; on Venus, probes designed to drop to the surface must deal with an environment that can melt metal.

So despite being the first planet to be visited by a probe, Earth's closest neighbour remains little-known. Venus's atmosphere contains a mystery substance, detected because it absorbs ultraviolet light, but so far unidentified. Scientists don't agree on how the planet's relatively young surface is remade, or how active its volcanoes are. The mechanism behind its enormous winds — which hit at several hundred kilometres per hour — is a mystery, as is why Venus rotates on its axis in the opposite direction to Earth. Does it have lightning? The jury is out.

Venus scientists feel that their planet is neglected. Despite a flurry of visits in the first decades of interplanetary exploration, NASA hasn't been to the planet since the Magellan mission ended in 1994. The European Space Agency's Venus Express orbiter filled a gap when it observed the planet from 2006 to 2014, but at €220 million (US$252 million) it was a relatively small mission, and it could only peer at Venus from orbit.

Now, after a rocky journey, Japan's Akatsuki mission — which many wrote off as lost when its main engine failed in 2010 — has entered Venusian orbit and is revealing intriguing results about the planet's climate (see page 157). In its wake is another glimmer of hope: two Venus projects are among five proposals shortlisted for NASA's next $500-million Discovery mission, launching in the early 2020s. VERITAS (Venus Emissivity, Radio Science, InSAR, Topography, and Spectroscopy) is a high-resolution radar mapper that would study the planet from the sky; the DAVINCI (Deep Atmosphere Venus Investigation of Noble gases, Chemistry, and Imaging) probe would sample the atmosphere during an hour-long plunge to the surface. Project leaders hope that compelling findings by Akatsuki will generate excitement about the planet at just the right time.

Given that life and the ability to sustain it will always be a selling point for an interplanetary mission, and that the only hope for life on Venus would be in its upper atmosphere, Venus's fall from favour might be understandable. But the planet holds a trump card. Increasingly, astronomers are searching for exo-Earths — extrasolar planets that, given their similarity to Earth, are a good bet for life. There, Venus can tell a cautionary tale. Despite starting out with all the ingredients for life, at some point Venus went rogue and became the hellish, acidic, dry planet it is today. Although life might not be found in a Venusian jungle, understanding why the planet took the path it did might be crucial to finding life elsewhere. ■

➲ NATURE.COM
To comment online, click on Editorials at:
go.nature.com/xhunqv

STEVEN PARRY DONALD, EDINBURGH

Origins of the obesity pandemic can be analysed

Statistical and biological methods are available to probe why the prevalence of obesity has risen more in some countries than in others, says **John Frank**.

What started the obesity pandemic? We remain unsure. And although we do not need to know the answer to tackle the symptoms, a clearer picture might produce better strategies.

Analytical methods for sorting out the epidemiological evidence on this question now lie within our reach. Economists have used these methods for many years to look at the impact of large natural experiments such as changes in policy. And in the past five years or so, epidemiologists have realized that they can be applied to health outcomes.

These statistical tricks include 'difference in differences' and 'fixed-effect variables' to control for 'unobserved heterogeneity' (confounding factors that were not measured and thus not taken into account). They are not foolproof, but they can be used to test — and often rule out — associations between population-level health changes and previous exposure to possible causes.

Obesity is particularly suited to this kind of epidemiological analysis because there is wide variation by country in when the problem began, and how quickly it developed.

As a Canadian living in Scotland (a severely overweight society), I have my own theory on the origins of obesity. And it could now be tested.

The most revealing 'epidemic curves' of the prevalence of obesity and overweight over time were published by the Organisation for Economic Co-operation and Development, with annual updates to 2012 (see go.nature.com/2bb5ns). They depict survey data on national obesity prevalence in nine developed countries and Mexico from the 1970s onwards. And, to this older epidemiologist's eye, there is a striking trend.

The United States, England, Australia, Mexico and possibly Canada show rapid growth in obesity prevalence to world-leading levels. Switzerland, Italy, Spain, South Korea and (at least initially) France, show slower growth, often with later onset, and certainly lower current levels. It is as if the two sets of countries had more, or less, resistance to the forces driving the pandemic. These are widely identified as some combination of increased calorie intake and unchanged or declining physical-activity levels.

What do the countries in each group have in common in relation to changes in lifestyle between 35 and 15 years ago? The societies that experienced later and slower weight gains have, in my opinion, much stronger cultural attachment to traditional cuisines, now thought to be healthier than most modern foods. (Mexico is obviously the exception; it is also in the economic backyard of the United States, where the pandemic hit first and hardest.) By contrast, the traditional foods of the predominantly English-speaking countries that saw an early and rapid rise in obesity are notoriously bland. When processed food laden with sugar, fat and salt arrived on the shelves, these peoples quickly switched.

THERE IS WIDE **VARIATION** BY COUNTRY IN **WHEN** THE PROBLEM BEGAN, AND HOW **QUICKLY** IT DEVELOPED.

To rigorously test this conjecture, what sorts of data could be analysed? Ironically, commercial sales data may hold more promise than human-health surveys. Self-reporting of diet and physical activity is well known to be unreliable: people forget, or say what they think they should say. Over extended periods, they may alter reports to fit evolving social norms, and different cultures may systematically provide differentially biased responses.

More promising are national time series of sales of those foods currently implicated in the pandemic's origins and spread, such as sugary beverages and fast foods with high caloric density, including French fries. Detailed sales data may be difficult to obtain, especially from big fast-food chains, but proxy measures, such as total sales, should be available. Indeed, publicly traded firms in this market have often boasted to their shareholders of historical sales increases.

Time series of credible data on physical activity at the national level are harder to come by, so proxy measures of sedentary habits, such as total hours of television-watching (and, more recently, 'screen time'), might be the best available.

This quasi-experimental approach could also test the most unusual hypothesis on the pandemic's origins. In 2013, infectious-disease researcher Lee Riley and his colleagues at the University of California, Berkeley, suggested that an increasing cumulative lifetime exposure to antibiotics could be responsible, driven by these powerful chemicals' presence in meat and dairy products in our food chain, and careless overprescribing in medical care.

To test this hypothesis, one could examine time series of sales of the relevant antibiotics, both human and veterinary. However, the lag times here are much less certain: it might take decades of antibiotic exposure, perhaps beginning in childhood or even at birth, to profoundly alter the bowel microbiota and presage weight gain.

A much better data source would be frozen stool samples obtained during national surveys, collected over the relevant time period. These could be analysed for effects (now well described) of prolonged antibiotic exposure on the bowel microbiota, whose nucleic-acid and protein fingerprints should still be detectable.

So there it is: a whole programme of new and potentially important scientific work, for anyone with the nerve — and resources — to execute it. Any takers? ■

➲ NATURE.COM
Discuss this article online at:
go.nature.com/echbzo

John Frank *holds a chair in public-health research at the University of Edinburgh, UK, and is professor emeritus at the University of Toronto, Canada.*
e-mail: john.frank@ed.ac.uk

RESEARCH HIGHLIGHTS

Selections from the scientific literature

GENETICS

Disease mutations but no disease

An analysis of genetic data from more than half a million people has uncovered 13 individuals who have disease-causing mutations but are healthy.

Mendelian diseases such as cystic fibrosis begin in childhood, can be caused by a single mutation and lack effective treatments. Rong Chen at the Icahn School of Medicine at Mount Sinai in New York City and his colleagues looked for mutations in 874 genes — linked to nearly 600 childhood genetic diseases — in roughly 589,000 people. They found 13 resilient people with mutations that usually cause 1 of 8 severe Mendelian diseases.

Further study of such individuals could lead to discoveries of gene variants that protect against disease, and could even lead to new treatment strategies, the authors say.

Nature Biotechnol. **http://dx.doi.org/10.1038/nbt.3514 (2016)**

ECOLOGY

Catfish face migration barriers

Amazonian catfish make the longest known freshwater migrations, covering thousands of kilometres, but their epic voyages are threatened by new dams.

Brachyplatystoma catfish can measure up to three metres in length, and are top predators. To study their migrations, Fabrice Duponchelle of the Institute of Research for Development in Montpellier, France, and his colleagues analysed the strontium isotope ratio in ear bones from 37 *Brachyplatystoma rousseauxii* captured near breeding areas in the Amazon basin. The authors found correlations between the strontium make-up of the bones and that of rocks in different parts of the river system. They suggest that young fish migrate downstream in the lower Amazon, then return upstream as adults, swimming some 8,000 kilometres to the area where they were hatched.

Two dams built recently on the Madeira River could prevent the fish from reaching their spawning grounds, which could have ripple effects through Amazonian food webs, the authors warn.

J. Appl. Ecol. **http://doi.org/bd45 (2016)**

JON WHITE/COLORADO GEOLOGICAL SURVEY

GEOLOGY

Fluid flow in landslides

Vibrations that ripple through rocks as they tumble downhill explain why some landslides travel farther than expected. The finding could help towns to better prepare for landslide hazards.

In 'long runout' landslides, falling rocks can move tens to hundreds of kilometres on flat land — more than ten times the height from which they fell. A team led by Brandon Johnson of Brown University in Providence, Rhode Island, modelled the forces in such landslides. The scientists found that vibrations caused by slides of sufficient size reduce the pressure between rock fragments, effectively lowering friction and allowing the rocks to flow like a fluid over long distances.

A similar effect could also occur along geological faults during earthquakes.

J. Geophys. Res. Earth Surf. **http://doi.org/bd4x (2016)**

ASTRONOMY

Black-hole disk launches jet

Scientists have caught one of the best glimpses yet of a jet of plasma streaming from the black hole at the heart of a distant galaxy.

Intense magnetic fields around black holes are thought to launch these beams, which travel nearly at the speed of light, but the beams' exact origins remain unknown.

Bia Boccardi of the Max Planck Institute for Radio Astronomy in Bonn, Germany, and her colleagues used a global array of radio telescopes to image the base of the jet at the core of the galaxy Cygnus A at high resolution. They found that the base is hundreds of times wider than the event horizon of the black hole, extending into the swirling disk of material that surrounds it.

This suggests that the rotation of the disk helps to launch the jet.

Astron. Astrophys. 588, **L9 (2016)**

ANTHROPOLOGY

War uncommon in prehistoric Japan

Hunter-gatherers living in Japan thousands of years ago were not particularly violent, adding weight to a contentious idea that violence and warfare were not the norm in early history.

Hisashi Nakao at Yamaguchi University in Japan and his colleagues analysed published data on the skeletal remains of hunter-gatherers from Japan's Jomon period, between 13,000 BC and 800 BC. The team calculated the percentage of skeletons showing evidence of fatal injuries from violence, and found that mortality from violence was low, averaging

1.8% over the entire Jomon period. Violent injuries were evenly distributed across the country, and the researchers found no hotspots of violence that might indicate warfare.

The findings are inconsistent with the idea that warfare is inherent in human nature, the authors say.

Biol. Lett. 12, **20160028 (2016)**

MICROBIOLOGY

Salmonella live on thanks to toxin

A toxin protein secreted by typhoid-causing bacteria seems to keep infected hosts alive, allowing the bacteria to persist in the body.

Salmonella enterica Typhi (*S.* Typhi), which causes typhoid fever in humans but not in mice, produces a DNA-damaging protein. To study this toxin's role in mouse infections, Teresa Frisan at the Karolinska Institute in Stockholm and her team engineered another strain of *S. enterica*, called *S.* Typhimurium (which causes illness in mice but does not normally make the typhoid toxin) to make the part of the toxin that damages host DNA.

Mice infected with the toxin-producing strain were less likely to become severely ill and had less gut inflammation than did mice infected with a control strain. Toxin-producing bacteria could still be found in the livers of mice six months after infection, in contrast to the control strain, which was undetectable in mice that survived the initial infection.

PLoS Pathog. 12, **e1005528 (2016)**

SCI. TRANSL. MED.

ASTROCHEMISTRY

Sugars made in simulated space

A key sugar found in DNA has been created in the laboratory under conditions similar to those around comets.

Ribose forms the backbone of DNA and RNA, but its ancient origin remains a mystery. Cornelia Meinert and Uwe Meierhenrich of the University of Nice Sophia Antipolis in France and their team created an artificial comet by condensing water, methanol and ammonia in a vacuum chamber at –195 °C. The material was irradiated with ultraviolet light to simulate the formation of cometary ices. The residues that formed when the material was warmed to room temperature contained ribose and other, similar sugars in amounts that were much greater than just trace levels.

The authors suggest that comets and meteorites are the source of organic molecules that made life possible on Earth.

Science 352, **208–212 (2016)**

HEART DISEASE

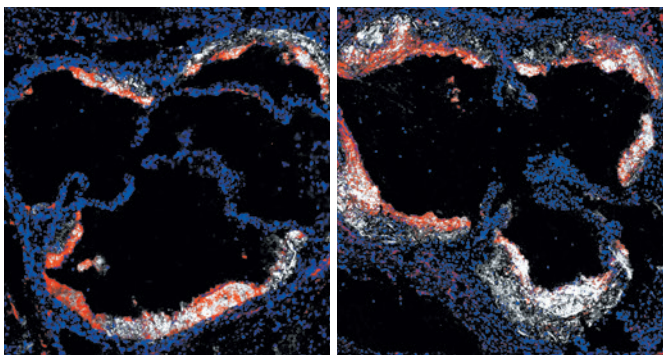
Molecule melts away cholesterol

The next weapon against heart disease could be a compound that is currently used to make drugs more soluble.

In atherosclerosis, plaques containing crystallized cholesterol clog up blood vessels. Eicke Latz of the University Hospital in Bonn, Germany, and his colleagues tested a compound called 2-hydroxypropyl-β-cyclodextrin, which increases the solubility of cholesterol, to see whether it reduced the plaques. They found that plaques shrank in atherosclerotic mice that had consumed cyclodextrin (blood vessel **pictured** left, cholesterol crystals in white), compared with plaques in the blood vessels of untreated animals (pictured right).

The drug bound to and dissolved the cholesterol crystals. It also increased cholesterol metabolism in immune cells called macrophages, which usually contribute to atherosclerosis by triggering inflammation in response to excess cholesterol. Cyclodextrin reprogrammed the cells in plaques, leading to increased transport of the dissolved cholesterol away from the plaques, and reducing harmful inflammation. Some of the same effects were seen in human plaque samples treated with the compound.

Sci. Transl. Med. 8, **333ra50 (2016)**



BIOCHEMISTRY

Bioplastic made from glucose

Researchers have combined three biochemical pathways to produce a biodegradable plastic from glucose in the laboratory.

Some industrial chemicals are made by microorganisms in bioreactors, but reengineering the organisms' metabolic pathways to boost yields is challenging, so researchers are keen to find cell-free production methods. Polyhydroxybutyrate bioplastic (PHB) can be made without cells, but the process requires expensive starting materials. To lower costs, James Bowie and his colleagues at the University of California, Los Angeles, devised a cell-free way to make PHB out of glucose. They designed a synthetic biochemical cycle comprising parts of three enzyme-driven pathways. Using two different concentrations of glucose, the team generated PHB at 86% and 94% yields. The yields and production rates were close to those required by industry.

With further improvements, this synthetic biochemistry approach could be used to produce chemicals at low cost, the authors say.

Nature Chem. Biol. **http://dx.doi.org/10.1038/nchembio.2062 (2016)**

NEUROIMMUNOLOGY

Protein linked to immune privilege

A protein found in neurons helps to limit inflammation in the central nervous system (CNS), contributing to the system's specialized immune environment.

The CNS can stave off excessive inflammation. This 'immune privilege' has been attributed to the blood–brain barrier that restricts the entry of certain immune cells, but recent work has suggested a role for other cells and molecules. Lieping Chen at Yale University in New Haven, Connecticut, and his colleagues found that SALM5, a protein involved in neuronal growth and development, inhibits inflammation in the mouse CNS. In animals with an autoimmune CNS disease, blocking a receptor for SALM5 or treating with an antibody against SALM5 aggravated symptoms. Applying SALM5 to certain immune cells in a lab dish suppressed their response to a pro-inflammatory molecule.

The findings could lead to treatments for inflammatory neurological diseases, the authors suggest.

Sci. Adv. 2, **e1500637 (2016)**

SEVEN DAYS *The news in brief*

EVENTS

Kepler scare

NASA mission managers were shocked to discover on 7 April that the exoplanet-hunting Kepler space telescope had entered emergency mode. Mission control was able to return it to normal operations three days later, but the cause of the malfunction remained a mystery as *Nature* went to press. This was the first software glitch in Kepler's seven years in space, although it previously suffered hardware breakdowns. The spacecraft has lost at least the first several days of a planet-hunting campaign that it was scheduled to begin on 7 April and conduct until 1 July. See go.nature.com/mu7woc for more.

China satellite lab

China has launched its largest-ever suite of microgravity and life-science experiments into orbit. The country's Shijian-10 probe left the Jiuquan Satellite Launch Center in Gansu province, northern China, on 7 April. It is carrying 19 experiments that include tests to assess the effects of radiation on genes as well as the influence of microgravity on materials, fluid physics and combustion. The early development of mouse embryos in microgravity will also be examined. After its 15-day mission, the bullet-shaped craft will re-enter Earth's atmosphere to be recovered from a landing site in Inner Mongolia.

Self-driving lorries

Six squads of automated lorries successfully arrived in Rotterdam in the Netherlands on 6 April after having driven themselves from Sweden, Belgium and Germany, with one fleet travelling more than 2,000 kilometres from Stockholm. The trial was part of the Dutch-government-led European Truck Platooning Challenge and included lorries from six different manufacturers. 'Truck platooning' involves two or more lorries connected by WiFi and driving in a convoy, with the first vehicle determining the speed and route. The technology aims to save fuel by enabling lorries to travel closer together, which reduces air drag.

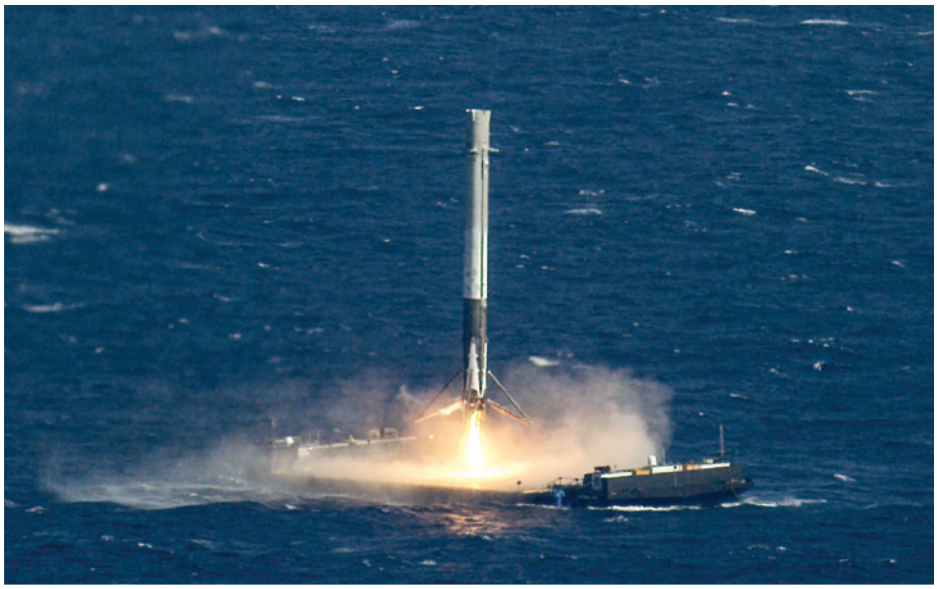
Bank climate plan

The World Bank announced a Climate Change Action Plan on 7 April to help countries to meet their commitments under the United Nations climate agreement signed in Paris in December 2015, and to prepare for unavoidable impacts of climate change. Under the plan, the bank will mobilize US$25 billion in private financing for clean energy by 2020. Among other actions, it will quadruple funding for clean transportation programmes and help to bring early-warning systems for natural disasters to 100 million people.

RESEARCH

Embryos edited

Researchers at Guangzhou Medical University in China have reported editing the genes of non-viable human embryos to try to make them resistant to HIV infection. The team collected a total of 213 fertilized human eggs, donated by 87 patients, that were unsuitable for implantation as part of *in vitro* fertility therapy because they contained an extra set of chromosomes. The researchers then used the CRISPR–Cas9 genome-editing technique to introduce into some of the embryos a mutation that cripples an immune-cell gene called *CCR5*. Some people naturally carry this mutation, which alters the CCR5 protein in a way that prevents the HIV virus from entering

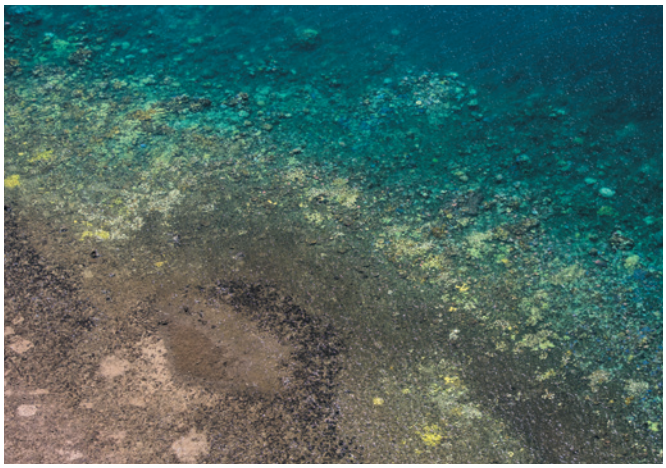


SPACE X

SpaceX rocket touches down at sea

SpaceX took a major step towards re-usable rockets when it flawlessly landed the first stage of its Falcon 9 rocket on an unmanned ship in the Atlantic Ocean, after an 8 April launch from Cape Canaveral, Florida. It was the first successful landing of the rocket at sea, following four attempts that resulted in crashes. The company, based in Hawthorne, California, returned an intact Falcon rocket to land in December last year. The latest flight delivered cargo to the International Space Station (ISS), including an expandable astronaut habitat designed by Bigelow Aerospace of North Las Vegas, Nevada. The previous SpaceX mission to the ISS failed when a Falcon 9 rocket broke apart after launch in June 2015.

the cells it tries to infect. Genetic analysis showed that 4 of 26 human embryos targeted were modified with the *CCR5* mutation. But in some embryos, not all sets of chromosomes harboured the mutation; some contained the unmodified gene, whereas others had acquired different mutations. In April 2015, a different China-based team announced that it had modified a gene linked to a blood disease in non-viable human embryos, igniting a worldwide storm of ethics concerns. See go.nature.com/igymgu for more.



ARC CENTRE OF EXCELLENCE FOR CORAL REEF STUDIES

ENVIRONMENT

Reef catastrophe

Huge swathes of coral in Australia's Great Barrier Reef are undergoing severe bleaching (**pictured**), according to aerial surveys. Many corals in the northern part of the reef are likely to die, because raised sea temperatures have caused them to expel the symbiotic algae that give them their colour. Researchers at the ARC Centre of Excellence for Coral Reef Studies in Townsville, Queensland, who are assessing the damage, say that more than 1,200 kilometres of the roughly 2,300-kilometre-long reef have bleached, and that the situation is substantially worse than in the two previous bleaching episodes in 1998 and 2002. See go.nature.com/ys7bau for more.

Cambodia tiger loss

Tigers are no longer breeding in Cambodia and the population there should be considered "functionally extinct", the conservation group WWF announced on 6 April in Phnom Penh. The last wild tiger there was seen on a camera trap in 2007 in the Mondulkiri Protected Forest. But the WWF noted that national estimates and data compiled by the International Union for Conservation of Nature suggest that global tiger populations have rebounded to 3,890, from about 3,200 in 2010. Cambodia plans to bring eight young tigers from India into its dry forests in the Eastern Plains by 2019, as part of the global Tx2 initiative aiming to double wild tiger populations by the year 2022.

BUSINESS

Pharma merger off

A marriage between two large pharmaceutical companies has been called off. Pfizer of New York City and Allergan of Dublin announced on 6 April that they had terminated a proposed merger process, which would have enabled the resulting company to take advantage of lower taxes in Ireland. The news came two days after the US Department of the Treasury unveiled stricter rules on companies that seek to move abroad to avoid US taxes. Pfizer pledged to announce by the end of the year whether it will spin off some parts of the company.

PEOPLE

NASA science chief

Former astronaut John Grunsfeld, who has overseen NASA's science portfolio since 2012, announced his retirement from the space agency on 5 April. The physicist and space-telescope expert flew five times on the space shuttle — including three visits to the Hubble Space Telescope — and was the lead spacewalker on the final flight to maintain and upgrade the telescope in 2009. As associate administrator for NASA's Science Mission Directorate, he was responsible for more than 100 missions, such as the New Horizons spacecraft that visited Pluto last year. Grunsfeld's deputy, Geoff Yoder, will take charge until a successor is chosen.

COMING UP

16–20 APRIL
The American Association for Cancer Research holds its annual meeting in New Orleans, Louisiana.
go.nature.com/q1t4fp

17–22 APRIL
The American Meteorological Society's 32nd meeting on hurricanes and tropical meteorology convenes in San Juan, Puerto Rico.
go.nature.com/pvszif

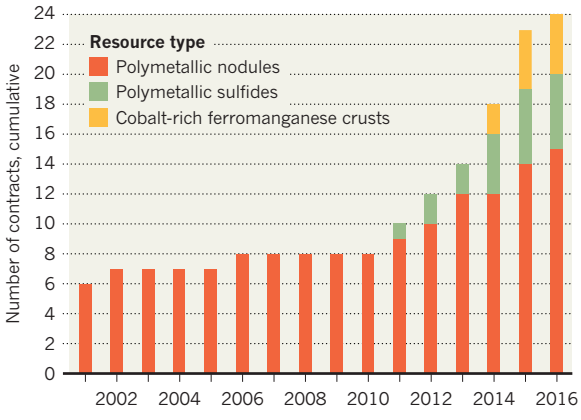
18–19 APRIL
London hosts the military space situational awareness conference.
go.nature.com/n6zeqm

TREND WATCH

Contracts with the International Seabed Authority (ISA), which regulates sea-bed mining in international waters, have picked up in recent years. Although commercial mining operations have not yet started, governments and corporations have signed contracts with the ISA to allow them to explore areas of the world's oceans for materials including manganese nodules, copper, zinc, cobalt and platinum. Researchers have warned about the environmental impacts, saying that stricter regulation is needed.

SEA-BED MINING HEATS UP

The number of contracts with the ISA shows growing interest in the exploration of mineral resources under the seas.



SOURCE: ISA

NATURE.COM
For daily news updates see:
www.nature.com/news

NEWS IN FOCUS

JOOST DE RAEYMAEKER/EPA

Locals wait in line for vaccinations in the Angolan capital Luanda after an outbreak of yellow fever.

PUBLIC HEALTH

Fears rise over yellow fever's next move

Scientists warn vaccine stocks would be overwhelmed in the event of large urban outbreaks.

BY DECLAN BUTLER

As the largest outbreak of yellow fever in almost 30 years continues to spread in Angola, scientists are warning that the world is ill-prepared for what would be a public-health calamity: the re-emergence of urban epidemics of the deadly infection, which could overwhelm vaccine stockpiles.

Yellow fever virus caused devastating outbreaks in cities in the past, but by the 1970s its mosquito carrier in urban areas — *Aedes aegypti* — had been wiped from large swathes of the globe; vaccination programmes also helped to confine the virus to the jungle. But now, as a result of the scaling-back of control efforts, *Aedes* mosquitoes have re-emerged in densely populated tropical and subtropical cities where many people are unvaccinated — and the Angolan situation has renewed fears that the virus might be poised to break out from the jungle.

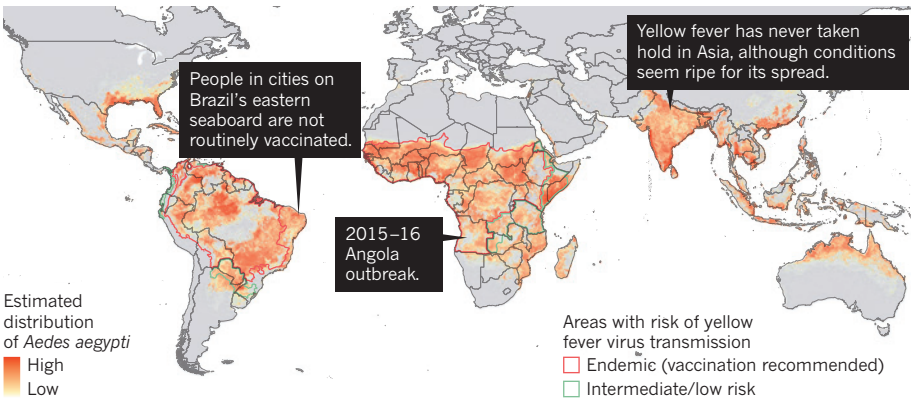
Worst of all would be if yellow fever gains a foothold in Asia — where, mysteriously, it has never become established despite apparently ideal ecological conditions. "We don't know if this will happen, but if it does it would be a public-health disaster," says Duane Gubler, a researcher in mosquito-borne diseases at Duke-NUS Medical School in Singapore.

Yellow fever, which is endemic in parts of South America and Africa, causes at least 60,000 deaths each year. Many people who become infected recover quickly (there are 84,000–170,000 annual infections, more than 90% of them in Africa), but some develop jaundice, bleed from their orifices and sustain fatal organ damage.

A 2006 initiative led by the World Health Organization (WHO) upped mass vaccinations and introduced routine immunization of children in many high-risk countries in Africa, ▶

WHERE MIGHT YELLOW FEVER GO NEXT?

An ongoing outbreak of yellow fever in Angola has scientists worried that the virus might spread to cities that harbour its urban carrier, the *Aedes aegypti* mosquito.



SOURCES: M. U. G. KRAEMER *ET AL. ELIFE* **4**, E08347; 2015 (MOSQUITO DISTRIBUTION); WWW.CDC.GOV/YELLOWFEVER/MAPS (TRANSMISSION RISK AREAS)

▶ but vaccination rates remain too low. The Angolan outbreak points to the continued risk posed by yellow fever: it began last December in the capital Luanda and has since spread to 6 of the country's 18 provinces. Officially, some 490 people have been infected and 198 have died, although the true figures are probably much higher.

The immediate concern is that the virus might spread to larger African urban centres, as happened in the biggest previous outbreak, which began in 1986 in Nigeria and ultimately infected 116,000 people and killed 24,000 (see 'Where might yellow fever go next?').

Africa's urban populations are now much larger than they were in the 1980s, notes Thomas Monath, a yellow fever and vaccine specialist who is chief scientific officer of NewLink Genetics in Ames, Iowa. People who contracted the infection in Angola have already carried it to Kenya, Mauritania and the Democratic Republic of the Congo, although this hasn't yet sparked new outbreaks.

FROM JUNGLE TO CITY

As long as the virus is confined mainly to small outbreaks in Africa, the world's vaccine production — just over 40 million doses annually — should be sufficient to replenish emergency stockpiles and contain outbreaks, says William Perea, who coordinates the WHO's Control of Epidemic Diseases Unit in Geneva, Switzerland.

But the fear is that yellow fever could follow the same path as other less-severe mosquito-borne diseases, such as dengue, chikungunya and Zika, which have already seen major urban epidemics tied to the resurgence of *Aedes* mosquitoes.

Scientists are struggling to assess that risk, in part because there is little research on the virus. In South America, for instance, despite endemic jungle yellow fever and *A. aegypti*-infested cities, urban outbreaks are almost unheard of. This may be because populations of monkeys and jungle mosquitoes (the reservoirs for the virus) are smaller than in Africa, and because of relatively high vaccination rates among people living in or near the jungle.

Yellow fever also seems to spread less easily by *Aedes* mosquitoes than do other viruses such as dengue, says Pedro Fernando da Costa Vasconcelos, an infectious-disease researcher who is director of the Evandro Chagas Institute in Ananindeua, Brazil.

Nevertheless, the WHO estimates that South America is now "at greater risk of urban epidemics than at any time in the past 50 years". Vaccination is officially recommended only in areas where the virus is endemic, because the vaccine can have serious — sometimes fatal — side effects in around 1 in 100,000 people. This means that few people are vaccinated on the densely populated eastern seaboard of Brazil, for example, because the region is the only part of the country where the virus is not endemic.

In Asia, the absence of yellow fever is an enigma: the continent has the monkeys, mosquitoes and climate in its warm regions that seem ideal conditions for the virus to thrive. Furthermore, infected travellers from elsewhere have introduced the virus to the region many times, and Asian populations don't have any specific resistance to yellow fever.

One hypothesis, says Gubler, is that strains of dengue and other related flaviviruses have for centuries been so prevalent that they offer cross-protection against yellow fever — and so viral loads are reduced to levels below those required to sustain mosquito-borne cycles of disease. But with the recent unbridled growth of large cities and mosquito-infested slums, Asia's past freedom from yellow fever may be no guide to its future, Gubler cautions.

The Angolan outbreak has also heightened concerns because hundreds of thousands of people from China and other Asian countries now work in Angola and in other at-risk parts of Africa, and many are unvaccinated, says Monath. Several unvaccinated individuals have fallen ill with yellow fever after returning from Angola to southern China.

VACCINE STOCKS

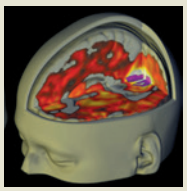
Many specialists want authorities to increase international vaccine stockpiles and to accelerate vaccination campaigns in endemic areas. But that would require boosting funding commitments to guarantee a market for manufacturers of the vaccine, Pereas says. Currently, there are only four suppliers worldwide, and supply is falling short of demand. (In a crisis, the WHO could dilute vaccines tenfold to boost its stocks without negating the vaccines' effectiveness. But the process would require a specific kind of single-use, small-volume syringe, which is not commercially available, Pereas adds.)

Gubler argues that vaccination should be considered in *A. aegypti*-infested cities close to endemic regions in Africa and South America. In Asia, the risk is too uncertain to recommend vaccination now, he says, but health-care workers there need to be trained to identify cases so that any outbreaks can be nipped in the bud. Renewed efforts to curb *A. aegypti* mosquitoes globally are also essential, he adds.

For now, the area of most immediate concern is Africa — where some countries, such as Nigeria, have less than 50% yellow fever vaccination coverage. The availability of vaccines has led to ill-founded complacency about the threat posed by yellow fever, warns Perea. "It's a neglected and forgotten disease." ■

MORE ONLINE

Q&A



Brain scans reveal how hallucinogens affect consciousness go.nature.com/kjbaue

MORE NEWS

- Door-to-door canvassing reduces transphobia go.nature.com/u7aaom
- HIV overcomes CRISPR gene-editing attack go.nature.com/rsnfsk
- Cosmological puzzle over Universe's expansion rate go.nature.com/wi18uv

NATURE PODCAST

A quantum video game, hearing voices and fault tolerance in the brain nature.com/nature/podcast

SPACE

Rescued Akatsuki spacecraft delivers first results from Venus

Streaked acidic clouds and a bow shape in the atmosphere are among its findings.

BY ELIZABETH GIBNEY

After an unplanned five-year detour, Japan's Venus probe, Akatsuki, has come back to life with a bang. On 4–8 April, the Japan Aerospace Exploration Agency (JAXA) presented the first scientific results from the spacecraft since it was rescued from an errant orbit around the Sun and rerouted to circle Venus, four months ago. These include a detailed shot of streaked, acidic clouds and a mysterious moving 'bow' shape in the planet's atmosphere.

Despite the probe's tumble around the Solar System, its instruments are working "almost perfectly", Akatsuki project manager Masato Nakamura, a planetary scientist at JAXA's Institute of Space and Astronautical Science in Sagamihara, Japan, announced at the International Venus Conference in Oxford, UK. And if another small manoeuvre in two years' time is successful, he said, the spacecraft might avoid Venus's solar-power-draining shadow, and so be able to orbit the planet for five years, rather than the two it was initially assigned.

Akatsuki, which means 'dawn' in Japanese, launched in 2010 and was supposed to enter into orbit around Venus later that year to study the planet's thick atmosphere. The mission would include looking for signs of active volcanos and other geology. But upon entry, a fault in a valve caused the probe's main engine to blow, and the craft instead entered an orbit around the Sun. As Akatsuki passed near Venus in December, JAXA engineers managed to salvage the mission by instructing the craft's much smaller, secondary thrusters to push it into a looping elliptical orbit around the planet. The results presented in Oxford were captured from this vantage point with a suite of five cameras that capture light ranging from infrared to ultraviolet.

A highly detailed shot of dense layers within Venus's sulfuric acid clouds elicited applause from the audience. The highest-quality infrared image of this view of Venus, it suggests that the processes that underlie cloud formation might be more complicated than thought, project scientist Takeshi Imamura told attendees.

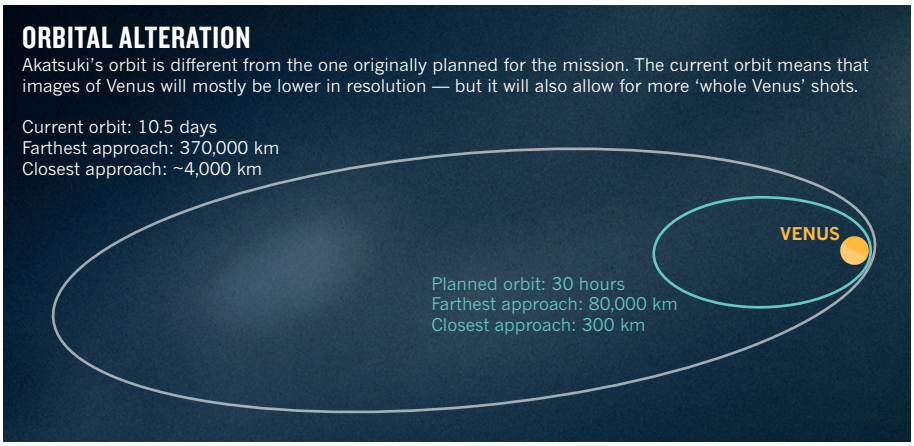
And the team expects still better results to come. The image was taken from 100,000 kilometres away — more than 10 times the probe's distance at its closest pass of Venus. "We will achieve better spatial resolution still," said Takehiko Satoh, principal investigator for the probe's 2-micrometre infrared camera, IR2, which took the image. "We promise to give a fantastic data set to the research community for years."

The bow shape, which was seen in thermal images taken using a long-wave infrared (LIR) camera, provided some intrigue. The moving cloud formation, which swept from pole to pole across the planet for days, seemed to rotate with Venus's surface, rather than with its much quicker-moving atmosphere.

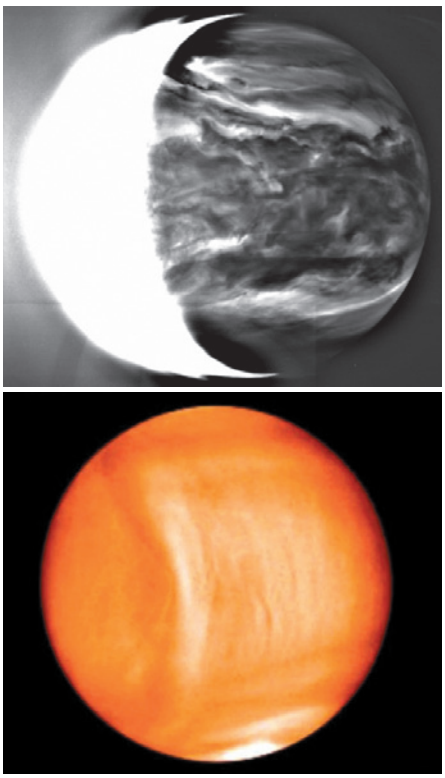
The motion suggests that the front could be linked to features on the ground, said Makoto Taguchi, who leads the LIR camera. Others at the conference were at a loss as to what may have caused it. "It's certainly mysterious," says planetary scientist Suzanne Smrekar of NASA's Jet Propulsion Laboratory in Pasadena, California.

Akatsuki's success has cheered researchers, especially because it is now the only working probe deployed at Venus. "The mood is very good," says Colin Wilson, a planetary scientist at the University of Oxford, UK. Akatsuki's orbit — which was tweaked slightly on 4 April to give the probe the best chance of lasting for years to come, as well as to provide a good scientific vantage point — will allow it to survey Venus's equator as originally planned. The resulting images will complement surveys of the planet's poles from the European Space Agency's Venus Express orbiter, which ended its mission in 2014.

But Akatsuki's new lease of life comes with compromises, too. Its current 10.5-day operational orbit takes it almost 5 times as far from Venus at its most distant point than its original intended orbit (see 'Orbital alteration'). Except for those taken during the short period when the probe sweeps close to the planet, images will be lower in resolution than planned. This means that studies that require detail, such as spotting flashes of lightning, will take longer. But the team said that it plans to make the best of the probe's wide orbit to take whole-Venus



SOURCE: JAXA



ISAS/JAXA

Streaked clouds and a mysterious 'bow' shape.

images that track large-scale features over time.

The mission has also not shrugged off all consequences of its long and unexpected cruise around the Sun. One camera malfunctioned in January, probably because of gradual contamination of a helium coolant with water vapour over the years, said Satoh. Engineers have now fixed the problem by warming the coolant to disperse the vapour, but it took a while. "We had a painful blank of about a month," says Satoh.

Planetary scientists outside of JAXA will have to wait a year from acquisition to access the data, but they are nonetheless excited by the probe's initial success. Two Venus-based projects are among five proposals shortlisted by NASA for possible launch in the early 2020s. The agency is expected to decide by the end of December, and Venus missions could get a boost from Akatsuki's success — especially if the orbiter finds intriguing features that require follow up, says Smrekar, who leads one of the Venus proposals that NASA is considering, the proposed VERITAS radar orbiter. "If they're able to see new volcanism, for example, it definitely makes the case for going back to explore more fully," she says. ■ SEE EDITORIAL P.148

BIOTECHNOLOGY

US rethinks crop regulation

Committee begins study to guide oversight of gene-edited organisms.

BY HEIDI LEDFORD

The industry that has blanketed more than 181 million hectares of the world's farmland with genetically modified (GM) crops is in the middle of a sea change. Improved techniques for altering crop genomes are already bringing a new generation of plant varieties to the market — and around the world, regulators are playing catch-up.

"A few brave countries have already made statements," says Piet van der Meer, a biologist and lawyer at Ghent University in Belgium. "But most are struggling with it."

On 18 April, the US National Academies of Sciences, Engineering and Medicine will begin its first meeting of a committee charged with ending the struggle. The committee, which is sponsored by the US Department of Agriculture (USDA) and two other agencies, has been asked to predict what advances will be made in biotechnology products over the next 5–10 years. It is scheduled to report by the end of the year on the steps that regulators need to take to prepare themselves. The result could inform an ongoing USDA effort to re-assess its process for evaluating engineered crops.

RITA ARIYOSHI/GETTY

The genetically engineered pink pineapple can be imported into the United States.

Researchers around the world are watching closely (see 'Global governance'). "Crops travel around the globe," says René Custers, manager of regulatory and responsible research at VIB, a life-sciences research institute in Ghent. "It is important to see what is happening in the rest of the world."

RIPE FOR CHANGE

Many feel that regulations in the United States, which grows more GM crops than any other country, are particularly ripe for change. The USDA itself has acknowledged that it might be over-regulating some crops if they have traits that have already been scrutinized. Also, it uses its authority to restrict the release of 'plant pests' as a way to regulate GM crops — an approach that applied widely in the 1980s, when crops were often created using genetic elements from plant viruses or bacteria.

But researchers have since developed tools that do not rely on these components. Over the past five years, the USDA has determined that about 30 types of GM plant — from soya beans whose oil has a longer shelf life, to pineapples with rose-coloured flesh — do not fall under its regulatory rubric. Some were made using gene-editing techniques.

"One of the things that has to happen is to plug

GLOBAL GOVERNANCE

Nations take a variety of approaches to regulating gene-edited products.

Like the United States, many countries are grappling with how to regulate crops that have been engineered using gene editing and other 'new breeding techniques' (NBTs).

Argentina In 2015, regulators decided that crops made using NBTs would be reviewed on a case-by-case basis.

Australia A 2013 workshop convened by the Food Standards of Australia and New Zealand recommended that NBT crops bearing simple deletions need not be considered GM food, but those with inserted genes should.

European Union The European Commission is expected this year to produce long-delayed advice on applying existing regulations to NBT crops.

Japan No official stance on gene-edited technologies, the products of which do not fall under the country's definition of a 'transgenic' crop.

Canada Decisions are made on the basis of whether the crops have new traits, irrespective of how the traits are produced.

New Zealand The Environmental Protection Authority determined that some crops made through NBTs would not be regulated, but the high court overruled the decision in 2014. H.L.

that huge hole," says Doug Gurian-Sherman, director of sustainable agriculture at the Center for Food Safety, an environmental-advocacy group in Washington DC. "Whether you think they're over-regulated or under-regulated or just not intelligently regulated, there's nobody who thinks this is appropriate."

And developers eager to market gene-edited varieties want clarity as to how the USDA will view the crops, says Daniel Voytas, chief science officer at Calyxt, a plant biotechnology company in New Brighton, Minnesota. The agency has already determined that it will not regulate several crops that have been developed using two editing tools — zinc-finger nucleases and TALENs — and it is currently considering a non-browning mushroom that was made using another, CRISPR–Cas9.

CASE BY CASE

These crops embody the simplest application of genome modification: deleting a small section of the genome to disrupt a gene. Calyxt, for example, used TALENs to edit a single gene in the parent plant and generate a variety of wheat with improved resistance to powdery mildew. On 11 February, the USDA informed Calyxt that it would not regulate the crop.

But more-sophisticated edits — such as rewriting genes or inserting new ones — are around the corner, Voytas says. "We don't understand how those crop varieties are going to be regulated," he says. "And they're already in the works."

On 5 February, the USDA released four broad regulatory scenarios that are open to public comment until 21 April. The draft proposed a definition of "products of biotechnology" that encompasses organisms in which segments of the genome have been deleted, added or altered. "Sometimes you are using these technologies to introduce genetic variation that already exists in wild relatives," says Custers. "The question is whether or not that differs from traditional plant breeding." Custers therefore advocates a definition that excludes plants carrying genetic changes that are already present in nature.

But including such plants in the definition does not mean that they would be heavily regulated, notes Greg Jaffe, director of biotechnology at the Center for Science in the Public Interest, a consumer advocacy group in Washington DC. "The USDA is capturing them under the rubric, but it sounds like they're also going to exempt many of them from oversight," he says.

Some activists are unlikely to support the idea. Gurian-Sherman notes that gene-editing technology is still relatively new, can be applied in many ways and sometimes makes unintended genetic changes. "We feel very strongly that this technology still needs to be regulated as we learn more about it," he says. "Maybe at some point it wouldn't need to, but this is still a new technology."■

FUNDING

Lab fights grant rejection and wins

Scientist hired lawyer to challenge European Commission.

BY EWEN CALLAWAY

Faced with a rejected grant application, scientists usually experience a range of emotions — shock, sadness, anger — before accepting the verdict and moving on. But when the European Commission rejected a €5-million (US$5.7-million) grant application from computational scientist Peter Coveney, he hired a lawyer and challenged the decision.

The successful appeal, made public on 29 March, highlights an aspect of the research-funding process that scientists rarely act on and almost never succeed at.

"I've been told by colleagues that you don't challenge the commission on anything," says Coveney, of University College London (UCL). "But if your research is in jeopardy as a part of poor decisions, then people should be prepared to challenge them."

Coveney thinks that his rare victory should encourage more researchers to appeal against decisions made by funders. But funding-agency administrators warn that the chances of success are low — and that fruitless appeals can waste time and resources. "If you're going to play the odds here, your chance of getting funded is substantially higher if you submit a revised proposal than if you go down the route of submitting an appeal," says Michael Lauer, director of the Office of Extramural Research at the US National Institutes of Health (NIH), the world's largest biomedical funder.

Appeals are uncommon in both Europe and the United States. Between 2007 and 2013, the European Commission's Framework Programme 7 received more than 106,000 grant applications, but although around 80% were rejected, only 3,683 decisions were appealed. Of these, 101 were re-evaluated and fewer than 10 succeeded in gaining funding. The US National Science Foundation, by comparison, received just 388 appeals between 2001 and 2014, 17 of which led to funding. Appeals at the NIH are similarly rare, says Lauer. Although the agency does not track them centrally, in eight years of overseeing cardiology-research grants, he saw just one successful challenge.

When the European Commission rejected the Coveney team's proposal to create a hub for applying computer modelling to biomedical and clinical data in May 2015, he was surprised. The 3-year project would involve 15 industrial and academic partners across Europe and use a consulting firm as project manager. Those elements fitted with a requirement for professional management, says Coveney, as outlined in the commission's funding call (part of a 7-year €78.6-billion programme called Horizon 2020). But he says that the reviews indicated that the team had brought in unnecessary partners by including the consulting firm, resulting in a poor score on that aspect.

FOLLOW THE RULES

Like some other funders, including the NIH, the commission has a formal 'redress' process that allows spurned scientists to ask for their grant applications to be re-reviewed. UCL advised Coveney that the odds of success were low. But he hired a law firm, Bindmans in London, to mount a challenge; his team incurred around £10,000 (US$14,000) in legal fees. He learned that his grant would be reconsidered in October 2015, and later that it had scored well enough on this subsequent review to be funded in February this year. He got official word of its approval last month. A representative of the commission confirmed that the grant's initial evaluation report contained incorrect information, leading to a new evaluation.

"If your research is in jeopardy as a part of poor decisions, then people should be prepared to challenge them."
Peter Coveney

"It is the only time I've challenged a grant decision so far in my life. I've seen a few dubious things happen in the past, but this one was so black and white," says Coveney. "It should send the message to people that they should think carefully and not just assume it's not worth it."

Not everyone agrees. "Lawyering up to ▶

get money is not something that strikes me as the way I'd do it," says Adrian Liston, an immunologist at the University of Leuven in Belgium. "I'd just take the grant to another agency."

Some researchers see Coveney's victory as an exception that proves the rule — science's version of 'You can't fight city hall'. Liston's own attempt to appeal a funding decision last year was foiled by a Kafkaesque process. When a funder that he declines to name denied a fellowship renewal for a postdoc in his lab, Liston was told that he first needed to request the reviews. They arrived two months later, and were positive. But the funder then told him that appeals had to be filed within a month of a rejection. "It's an appeals process on paper, but they make it so it can't ever be used," he says.

DIFFERING OPINIONS

A lack of expertise on the review panel is one of the few grounds on which the NIH says that it will grant an appeal, in addition to factual errors, bias or conflicts of interest on the part of reviewers. But Lauer says that such complaints often boil down to differences of opinion, which can't be appealed against.

Researchers are personally invested in their grant proposals, making rejection that much harder to handle, says Sally Rockey, Lauer's predecessor at the NIH, who is now executive director of the Foundation for Food and Agriculture Research in Washington DC. "People have a tough time separating their emotions from the actual review itself."

There may now be more motivation than ever to appeal against grant rejections, because the success rates of grant applications are in decline at many funding agencies, notes Björn Brembs, a neurobiologist at the University of Regensburg in Germany who still bemoans the denial in 2003 of a grant extension that he requested in from Germany's major funding agency, the DFG. "At a certain threshold of desperation and lack of alternatives, then an appeal doesn't seem as much of a cost any more," he says.

Appeals could waste the time of overworked agencies already faced with far too many strong applications to fund, warns Douglas Kell, a biologist at the University of Manchester, UK, and former head of the country's Biotechnology and Biological Sciences Research Council. Like the DFG, as well as Britain's other government funders, the biotechnology council does not have a formal appeals process.

"There are lots of things I would say we could do to improve funding procedures," says Kell. "But letting people bitch about the ones that go down isn't one of them." ■ SEE EDITORIAL P.147

SCIENCEATHOME.ORG

Games enable researchers to appeal to the public for help in solving scientific problems.

PHYSICS

Quantum world may be intuitive

A computer game suggests that the human mind is adept at grasping the bizarre laws of quantum mechanics.

BY ELIZABETH GIBNEY

With particles that can exist in two places at once, the quantum world is often considered to be inherently counterintuitive. Now, a group of scientists has created a video game that follows the laws of quantum mechanics, but at which non-physicist human players excel (J. J. W. H. Sørensen *et al. Nature* **532,** 210–213; 2016).

One implication of the team's results is that efforts to use computer games to crowdsource solutions to science problems can now be extended to quantum physics (see page 184). In the past, such gamification projects have been limited to challenging but less mind-bending problems, such as protein folding.

But the work also suggests that the human mind might be more capable of grasping the rules of the bizarre quantum world than previously thought — a revelation that could have implications for how scientists approach quantum physics, says Jacob Sherson, a quantum physicist at Aarhus University, Denmark, who led the study. "Maybe we should allow some of that normal intuition to enter our problem solving," he says. Scientists studying quantum foundations have also long said that finding a more intuitive approach to quantum physics could help to crack outstanding puzzles, although many doubted that this would ever be possible without new theories.

The game, called *Quantum Moves*, is based on a real problem in quantum computing: how fast a laser can move an atom between wells in an egg-box-like structure without changing the energy of the atom, which is in a delicate quantum state. In the quantum world, speed and energy are a trade-off limited by Heisenberg's uncertainty principle, so the trick is to find the sweet spot where the transition from one place to another is as fast as possible without disturbing the quantum state. Endless possible combinations of movement and timing exist, and scientists have designed computer algorithms to try to solve the problem.

In the game, an atom is represented by what looks like a liquid sloshing around in a well, which reflects the wave-like nature of a quantum particle. In one level, players move

a cursor to control a second well, which they use to collect the sloshing liquid and take it back to a base. The liquid behaves according to the laws of quantum mechanics rather than like an actual bucket of water — for example, to pick up the liquid, players can get it to 'quantum tunnel' from one well to another, something that players must learn to adapt to. Once they find ways to transfer the liquid, a computer can then convert their mouse movements to solutions to the real-world quantum egg box.

Sherson's team got around 300 people to play this level a total of 12,000 times on a volunteer-research platform called ScienceAtHome. The researchers then fed the human solutions into a computer for further refinement. Not only were more than half of the human-inspired solutions more efficient than those produced by just computer algorithms, but the two best hybrid strategies were faster than what the quickest computers had been able to achieve working alone. "I was completely amazed when we saw the results," says Sherson.

HUMAN ADVANTAGE

What abilities humans bring to the mix is unclear. Although an interest in physics seems to correlate with ability in the game, success did not correlate with years spent studying quantum physics. Sherson suggests that the superior human strategies stem from the mind's ability to capture the essence of a problem. Quantum concepts may seem less bizarre to people in a game than they do in other contexts, because it is an environment in which they expect rules to be broken, adds Sabrina Maniscalco of the Turku Centre for Quantum Physics in Finland, who runs an event aimed at making games that might benefit quantum physics.

"We should try to be more spontaneous and intuitive about problem solving."

To Sherson, the results also suggest that physicists could use their own intuition more. "We should try to be more spontaneous and intuitive about problem solving," he says. To that end, his team is building a version of the game in which physicists can tweak the scenario to represent different set-ups, potentially offering them new insights into their work.

Other quantum physicists agree that the finding that people can develop an intuition for quantum processes is surprising, but think that scientists already use intuition to solve quantum problems, at least at the mathematical level. By playing the game, people perhaps gain a form of that intuition, says Seth Lloyd, a physicist at the Massachusetts Institute of Technology in Cambridge. He notes that before babies learn to expect an object to stay where it is, they have a form of quantum intuition, which they lose. "Before three months, if it disappears, they guess that's just how things are in the world. After three months, they think, 'Where'd the toy go'?"

Lloyd also says that much of the success of *Quantum Moves* is due to its clever design, which successfully translates a quantum problem to a visual one, but which could fail with more-complex quantum problems.

Physicists who are trying to develop quantum-computing algorithms already play around with graphical interfaces to help them to improve on existing solutions, says Charles Tahan, a theoretical physicist at the University of Maryland in College Park.

But Tahan does think that teaching quantum intuition through games has benefits. He has developed another game, *Meqanic*, that gets players to perform basic quantum computations and intuit the rules as they play. He hopes that it could boost student's abilities and help to find individuals who have an untapped natural flair for the field. ■

CORRECTION

In the News story 'Controversial dark-matter claim faces ultimate test' (*Nature* **532,** 14–15; 2016), the last paragraph was amended to better reflect Katherine Freese's views on the DAMA collaboration's results.

THE PERFECT BLEND

The next frontier in cancer immunotherapy lies in combining it with other treatments. Scientists are trying to get the mix just right.

BY HEIDI LEDFORD

In cancer research, no success is more revered than the huge reduction in deaths from childhood leukaemia. From the 1960s to the 2000s, researchers boosted the number of children who survived acute lymphoblastic leukaemia from roughly 1 in 10 to around 9 in 10.

What is sometimes overlooked, however, is that these dramatic gains against the most common form of childhood cancer were made not through the invention of new drugs or technologies, but rather through a reassessment of the tools in hand: a dogged analysis of the relative gains from different medicines and careful strategizing over how best to apply them side by side as combination therapies.

"It wasn't just about pounding drugs together," says Jedd Wolchok, a medical oncologist at Memorial Sloan Kettering Cancer Center in New York City. "It was about understanding the mechanism and figuring out what should be given when."

That lesson has particular relevance in cancer research today. A new class of immunotherapies — which turn the body's immune system against cancerous cells — is elevating hopes about combination therapies again. The drugs, called checkpoint inhibitors, have already generated great excitement in medicine when applied on their own. Now there are scores of trials mixing these immune-boosting drugs with one another, with radiation, with chemotherapies, with cancer-fighting viruses, with cell treatments and more. "The field is exploding," says Crystal Mackall, who leads the paediatric cancer immunotherapy programme at Stanford University in California.

Fast-moving trends in cancer biology often fail to meet expectations, and little is yet known about how these drugs work together. Some observers warn that the combinations being tested are simply marriages of convenience — making use of readily available compounds or capitalizing on business alliances. "In many cases, we're moving forward without a rationale," says Alfred Zippelius, an oncologist at the University of Basel in Switzerland. "I suspect we'll see some disappointment in the next few years with respect to immunotherapy."

But many clinicians argue that delay is not an option as their patients queue up for the next available clinical trial. "Right now I have more patients that could benefit from combinations than there are combinations being tested," says Antoni Ribas, an oncologist at the University of California, Los Angeles. "We're always waiting on the next slot."

LYING IN WAIT

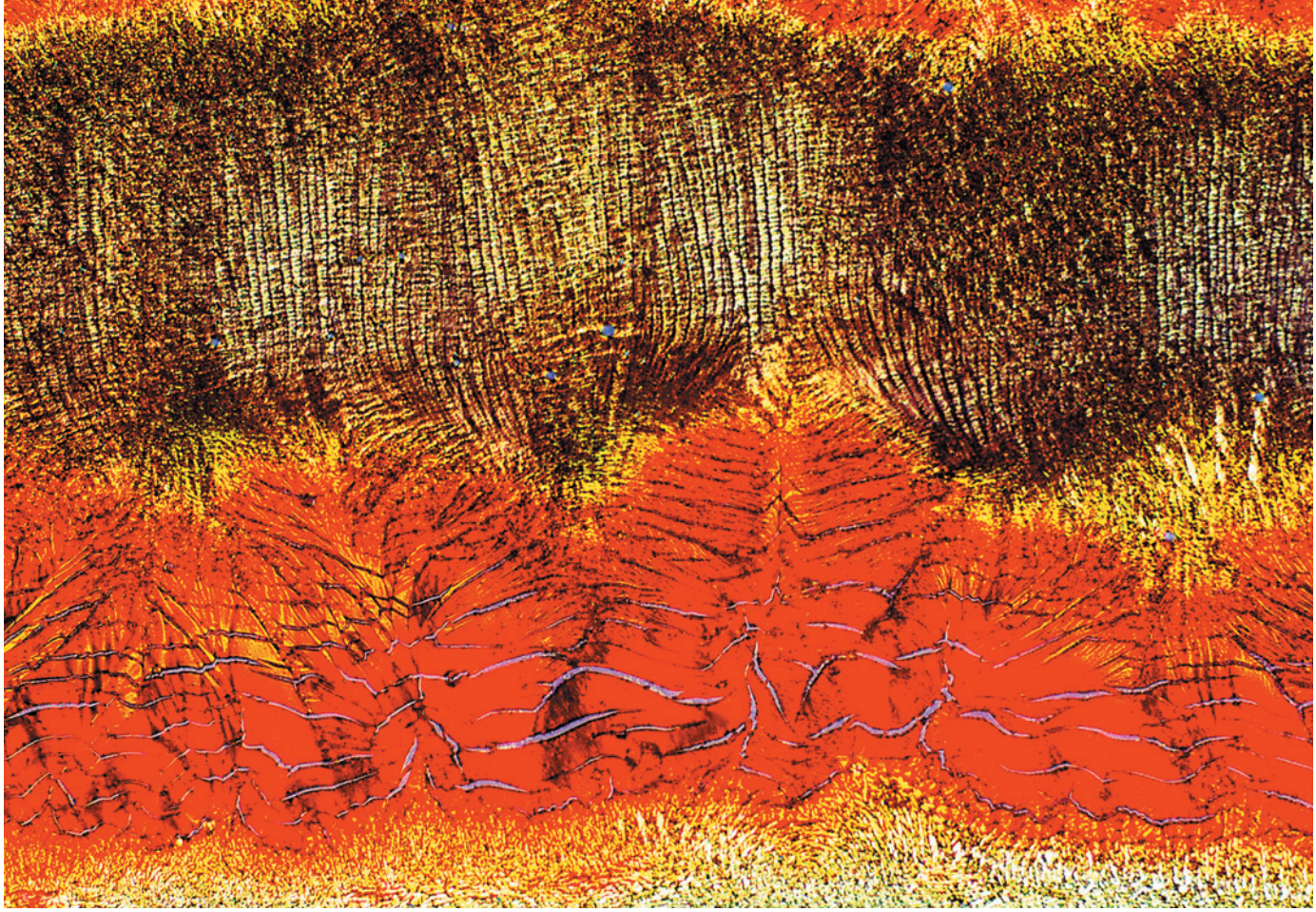
Immunotherapies have been more than a century in the making, starting when physicians first noticed mysterious remissions in a few people with cancer who contracted a bacterial infection. The observations led to a hypothesis: perhaps the immune system is able to kill tumours when made hypervigilant by an infection. The concept has vast appeal. What better way to beat a fast-evolving biological system such as a tumour (see page 166) than with a fast-evolving biological immune system? But it took decades for researchers to turn that observation into something useful.

Part of the trouble, they eventually learned, is that tumours suppress the immune response. T cells, the immune system's weapon of choice against cancer, would sometimes gather at the edge of a tumour and then just stop.

It turned out that a class of molecules called inhibitory checkpoint proteins was holding those T cells at bay. These proteins normally protect the human body from unwarranted attack and autoimmunity, but they were also limiting the immune system's ability to detect and fight tumours.

In 1996, immunologist James Allison, now at the University of Texas MD Anderson Cancer Center in Houston, showed that switching off a checkpoint protein called CTLA-4 helped mice to fend off tumours[1]. The discovery suggested that there was a way to re-mobilize T cells and beat cancer.

In 2011, the US Food and Drug Administration (FDA) approved the first checkpoint inhibitor — a drug, called ipilimumab, that inhibits CTLA-4 — to treat advanced melanoma. The improvements were modest: about 20% of patients benefited from ipilimumab, and the survival gain was less than four months on average[2]. But a handful of recipients are still alive a decade after starting the therapy — a stark contrast with most new cancer drugs, which often benefit more patients in the short term, but don't have a durable response (see 'Desperately seeking survival').



MARGARET OECHSLI

Some cancer drugs (pictured here, dried adriamycin viewed under a microscope) might work better when paired with immunotherapies.

Ipilimumab was at the leading edge of a flood of checkpoint inhibitors to enter clinical trials. The drug's developer, Bristol-Myers Squibb of New York, followed up with the approval of nivolumab, which inhibits the protein PD-1. And a host of other companies have jumped into the immunotherapy fray, as have academics such as Edward Garon at the University of California, Los Angeles. "Our group gladly shifted into this," says Garon, who began focusing on checkpoint inhibitors in 2012. "It was very clear this was going to have a major impact."

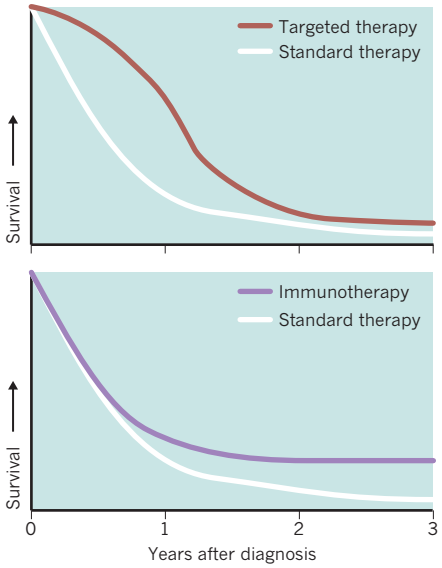
But even as the family of checkpoint inhibitors was rapidly expanding, the drugs were running up against the same frustrating wall: only a minority of patients experienced long-lasting remission. And some cancers — such as prostate and pancreatic — responded poorly, if at all, to the drugs.

Further research revealed a possible explanation: many people who were not responding well to the drugs were starting the treatment without that phalanx of T cells waiting at the margins of their tumours. (In the lingo of the field, their tumours were not inflamed.) Researchers reasoned that if they could raise this T-cell response first, and recruit the cells to the edges of the tumour, they might get a better result with the checkpoint inhibitors.

That realization fuelled a rush to test combinations of drugs (see 'Combinatorial explosion'). Radiation and some chemotherapies kill enough tumour cells to release proteins that the immune system might then recognize as foreign and attack. Vaccines containing these proteins, called antigens, could have a similar effect. "On some level, one can make an argument for almost any drug combining well with an immunotherapy," says Garon. "And obviously we know not all of them will."

DESPERATELY SEEKING SURVIVAL

Patients generally respond well to targeted therapies (top), which are directed at specific mutations in a cancer, but only for a short time. Checkpoint immunotherapies (bottom) do not help as many people, but those they do help tend to live longer. Oncologists are trying to get the best out of both strategies by combining the drugs.



MIXING IT UP

One of the first combinations to be tested was made up of two immunotherapies — ipilimumab and nivolumab — at once. Although the targets of these drugs both do the same job, silencing T cells, they do so in different ways: CTLA-4 prevents the activation of T cells; PD-1 blocks the cells once they have infiltrated the tumour and its environment. And treating mice with compounds that block both proteins yielded a more-inflamed tumour as well[3]. "There was reason to think that if you block both, the T cells will be even more ready to kill the tumours," says Michael Postow, an oncologist at Memorial Sloan Kettering.

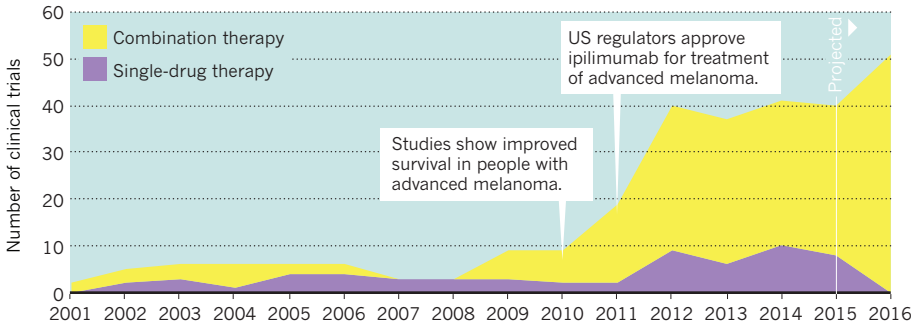
Together, ipilimumab and nivolumab boost response rates in people with advanced melanoma from 19% with just ipilimumab to 58% with the combination[4]. The combination also produces more-dangerous side effects than using either drug alone, but physicians are learning how to treat immunotherapy reactions, says Postow.

Ipilimumab generally doesn't help people with lung cancer when given on its own, but researchers are now testing it with nivolumab. Normally, they would not have bothered to investigate a combination involving a drug that had failed on its own, Garon says.

The new approach is grounded in immunology, but some researchers worry that

COMBINATORIAL EXPLOSION

Ipilimumab, the first approved checkpoint inhibitor, has been tested in dozens of clinical trials since 2001. And like many other drugs in its class, it is increasingly being tested in combination with other therapies.



SOURCE: CLINICALTRIALS.GOV

the effort could be wasted, he adds. Researchers are also testing inhibitors of other checkpoint proteins, including TIM-3 and LAG-3, in combination with those that block PD-1.

The combination approach is breathing life into drugs that had been shelved. For example, a protein called CD40 stimulates immune responses and has shown promise against cancer in animals. But in the wake of disappointing early clinical trials, some companies put their CD40 drugs to the side.

Years later, mouse studies showed that combining CD40 drugs with a checkpoint inhibitor could boost their effect. Now, at least seven companies are developing them. Cancer immunologists have listed the protein as one of the targets they are most interested in studying, says Mac Cheever, a cancer immunologist at the Fred Hutchinson Cancer Research Center in Seattle, Washington.

Cancer vaccines — long pursued by researchers but burdened by repeated failures in clinical trials — may also see a renaissance. There are now more than two dozen trials of cancer vaccines that make use of a checkpoint inhibitor.

Some promising combinations have been uncovered by serendipitous clinical observations. Researchers at Johns Hopkins University in Baltimore, Maryland, were conducting trials of epigenetic drugs, which alter the chemical tags on chromosomes. They shifted a handful of people with lung cancer who had not responded to the drugs to a clinical trial of nivolumab. Five of them responded — a much higher proportion than expected. The discovery became the seed for an ongoing clinical trial launched in 2013 to study combinations of epigenetic drugs and immunotherapies. Preclinical work has now provided evidence that epigenetic drugs can affect aspects of the immune response.

RIDING THE WAVE

These chance observations could lead to real advances, says Wolchok. "We're riding the wave of enthusiasm." But extracting the most from these combinations will require more well-designed preclinical studies to support the human ones. Just as attention to combinations of chemotherapies fuelled advances in treating paediatric leukaemias, the current combinatorial craze will require careful planning to work out the right pairings and timing of therapies.

Another class of drug, known as targeted therapies, could also receive a significant boost from immunotherapy. These drugs, which target proteins bearing specific mutations, generate a high response rate when given to patients with those mutations, but the tumours often develop resistance to the drugs and come roaring back. Coupling targeted therapies with a checkpoint inhibitor, researchers reason, could yield both high response rates and durable remissions.

One of the first targeted therapies for melanoma was an inhibitor that is specific to certain mutations in BRAF proteins that can drive tumour growth. However, an early attempt to combine this drug with ipilimumab was aborted when trial participants showed signs of possible liver damage[5]. No one was injured, but for some it was an important reminder that combinations can yield unanticipated side effects. "It was a good lesson for us to learn," says Wolchok. "It will not be as simple as we imagined."

Paying careful attention to sample collection during clinical trials would help researchers to catch toxicity problems early, says Jennifer Wargo, a cancer researcher at MD Anderson. "We're making mistakes by looking just at clinical endpoints," she adds. "We need to be smarter about how we run these trials."

In one of his latest trials, Wolchok wants to combine immunotherapy with a drug that targets a cellular pathway that some cancer cells use to maintain their rapid division. Cancers with mutations in this pathway, which is regulated by the protein MEK, can be extraordinarily difficult to treat.

But the pathway is also important for T-cell development, so Wolchok is working to determine the right timing for the treatment. One approach could be to use a MEK inhibitor to quiet tumours in mice and to release tumour antigens. He would then wait for the T-cell response to rejuvenate before adding the immunotherapy. "You want to make sure you're not trying to activate the immune system at the same time you're turning off that signalling," he says.

Garon is watching such trials with optimism, but he's aware that there may be a limit to how well combinations will perform. He sees a cautionary tale in a drug from an earlier era that works mainly in people with a mutation in the protein EGFR. Researchers spent a decade trying to find drugs that could turn a non-responding patient into a responder. "It is now clear that there probably is no such agent," he says. "I'm hopeful we won't be repeating that same response, but we have to watch our data cautiously."

DATA FRENZY

Researchers are so ravenous for those data that the results are being unveiled at major meetings at an earlier stage than in the past, he adds. "People are getting up and presenting response rates when the number treated is five," Garon says. "We generally have had a higher threshold than that." He worries that presenting such early data could prompt community physicians in the audience to start making decisions on treatments before they are appropriately studied.

The excitement is also fuelling a frenzy of clinical trials that are often based on speed rather than rationale. "Right now I'm kidding myself if I say I'm picking a combination because I have a scientific reason to pick it," says Mackall. "It's likely to just be what was available."

The strategy may still produce some wins. "There is plenty of opportunity for serendipity now," says Robert Vonderheide, who studies CD40 at the University of Pennsylvania in Philadelphia. But as the field matures, he says, this could give way to a more-systematic approach, similar to the careful planning and testing of variables used for paediatric leukaemias.

Despite his concerns, Garon is excited to be a part of the immunotherapy wave. Last autumn, he and his colleagues held a banquet for the patients who had been enrolled in his first immunotherapy trials three years earlier. These were the lucky survivors — the few who had shown a dramatic response. As he looked around the table at the guests of honour, he marvelled at their recovery. All had been diagnosed with advanced lung cancer, and many had been too weak to work. Now they were talking about their families, re-embarking on careers and taking up old hobbies such as golf and running. "We've never been able to hold a banquet like that before," he says. "I would love to hold many more." ■

Heidi Ledford *writes for* Nature *from Cambridge, Massachusetts.*

1. Leach, D. R., Krummel, M. F. & Allison, J. P. *Science* **271,** 1734–1736 (1996).
2. Hodi, F. S. *et al. N. Engl. J. Med.* **363,** 711–723 (2010).
3. Curran, M. A., Montalvo, W., Yagita, H. & Allison, J. P. *Proc. Natl Acad. Sci. USA* **107,** 4275–4280 (2010).
4. Larkin, J. *et al. N. Engl. J. Med.* **373,** 23–34 (2015).
5. Ribas, A., Hodi, F. S., Callahan, M., Konto, C. & Wolchok, J. *N. Engl. J. Med.* **368,** 1365–1366 (2013).

Tumour cells can evolve resistance to chemotherapy drugs such as oxaliplatin, shown under a microscope.

MARGARET OECHSLI

CANCER: AN EVOLVING THREAT

Tumours are subject to the same rules of natural selection as any other living thing. Clinicians are now putting that knowledge to use.

BY CASSANDRA WILLYARD

About six years ago, Alberto Bardelli fell into a scientific slump. A cancer biologist at the University of Turin in Italy, he had been studying targeted therapies — drugs tailored to the mutations that drive the growth of a tumour. The strategy seemed promising, and some patients started to make dazzling recoveries. But then, inevitably, their tumours became resistant to the drugs. Time and time again, Bardelli would see them relapse. "I stumbled into a wall," he says. The problem wasn't the specific mutations, Bardelli realized: it was evolution itself. "Unfortunately, we are facing one of the most powerful forces on this planet," he says.

Researchers have long understood that tumours evolve. As they grow, mutations arise and populations of genetically distinct cells emerge. The cells that are resistant to treatment survive and expand. No matter what medication physicians apply, it seems, the tumour adapts. And it has been difficult for researchers to unpick this process, because cancer evolves inside the body over the course of years. "We used to say to patients all the time that cancers are evolving in a Darwinian manner, but we didn't have a huge amount of evidence at our disposal to really formally prove that," says Charles Swanton, a cancer researcher at the Francis Crick Institute in London.

That is beginning to change. Thanks to advances in sequencing technology and the development of massive collections of samples and clinical data, scientists are piecing together a more precise picture of how cancer evolves, revealing the roots of resistance and, in some cases, finding out how it might be overcome. With a growing arsenal of treatments, biologists are trying to capitalize on these insights.

"Cancer is continuously adapting, therefore we have to do so as well," Bardelli says. In that spirit, last year he shifted the focus of his lab to studying the evolution of cancer. His

team has modelled[1] how colorectal cancers respond to targeted therapies that are given in combinations, potentially revealing ways to prevent the tumour cells from becoming resistant. "We have very exciting data now on the possibility to track and treat evolution," he says.

TREE OF LIFE

Cancer cells harbour a staggering array of mutations. In 2012, when Swanton and his colleagues sequenced multiple biopsies from two people with kidney cancer, they found that even within a single person, no two samples were the same[2]. The team examined not only the primary tumour, but also the satellite tumours — called metastases — that had spread throughout the patients' bodies. In each person, the team found more than 100 mutations in the various tumour samples analysed; only about one-third of them occurred in all samples.

The relationships between the various cancerous cells from a single person can be plotted out in much the same way as evolutionary biologists plot relationships between species: by drawing phylogenetic trees, branching diagrams that trace 'descendants' back to a common ancestor. Mutations that occur in the first malignant cells, those in the trunk of this evolutionary tree, will end up in all the tumour cells; mutations that arise later will be found only in the tree's branches. To eliminate the tumour, Swanton says, one must attack the mutations in the trunk.

Therapies that target some of these trunk mutations already exist, and they often produce dramatic responses at first. But then resistance develops, as Bardelli found. "We're so fixated on 'the smaller the tumour gets, the better', but what one doesn't think about is what one has left behind," Swanton says. "You're often leaving resistant clones that you can't treat." But he thinks that by targeting multiple trunk mutations at the same time, researchers might have a shot at wiping out the cancer. Chances are slim that a single cancer cell would be able to evade a two- or three-pronged attack.

One way to do this is to use combinations of targeted therapies. "In theory, they can work," says Bert Vogelstein, a cancer geneticist at the Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins University in Baltimore, Maryland. In fact, when he and evolutionary biologist Martin Nowak at Harvard University in Cambridge, Massachusetts, modelled the strategy, they found that two targeted medicines for which no common resistance mechanism exists would be enough to control metastatic cancer[3]. For people with a large number of metastases, the model suggested that three therapies would be needed.

Researchers are already beginning to test combinations of targeted therapies in the clinic. However, Swanton points out that there are no targeted drugs for the vast majority of mutations. And combining existing drugs in a way that won't harm the patient has proved tricky. So Swanton is focusing on immunotherapies — strategies that help the immune system to recognize and destroy cancer cells (see page 162).

The immune system identifies threats, in part, by surveying the surfaces of cells, looking for molecules called antigens that can signal trouble within. The genetic defects in the DNA of a cancer cell can sometimes encode antigens that will trigger an immune response. But Swanton and his colleagues wondered whether it matters if the immune system responds to an antigen that arises from the cancer's evolutionary trunk or to one that springs from its branches.

> "CANCER IS CONTINUOUSLY ADAPTING, THEREFORE WE HAVE TO DO SO AS WELL."

In a paper published in March[4], he and his colleagues examined samples from the Cancer Genome Atlas, a collection of genetic and clinical data from thousands of people with cancer. They found that people with lung cancer who had lots of trunk antigens — and a high proportion of trunk antigens to branch antigens — survived longer than those who had either few trunk antigens or a higher proportion of branch antigens. What's more, people with many trunk antigens seemed to respond better to immune therapies. That makes sense, Swanton says, because if the immune system targets trunk antigens, it's hitting most of the cancer cells, rather than "nipping off little branches".

The research is still in its infancy, but Swanton is leading a clinical study that could help to confirm his findings. The study, called TRACERx — Tracking Cancer Evolution through Treatment (Rx) — will follow 850 people from lung-cancer diagnosis through treatment and, in some cases, until death. It will document genetic changes in their tumours over time to examine how lung cancer evolves, and how treatment influences that process.

Once he has the data, Swanton hopes to raise enough money to test treatment strategies based on evolution. One approach would be to identify immune cells in a tumour, grow them in a lab, and then infuse them back into the patient — a technique called adoptive cell transfer. Similar strategies already in use select immune cells that recognize any cancer antigen, but Swanton's group would select those that are primed to recognize the trunk antigens that occur on all cancer cells.

This strategy would not be cheap, but neither is doling out a string of targeted therapies only to watch them all eventually fail. "Each course of therapy costs between US$10,000 and $100,000," Swanton says. If researchers could develop a therapy that would cure metastatic cancer, "the whole cost–benefit analysis and the health economic models change dramatically".

CELLULAR COMPETITORS

Applying evolutionary principles might help the immune system to vanquish tumours. Robert Gatenby, a molecular oncologist at Moffitt Cancer Center in Tampa, Florida, has a more modest goal: he hopes to help people to live with their disease. Gatenby began thinking about cancer as an evolutionary problem in the early 1990s, when he was working at Fox Chase Cancer Center in Philadelphia, Pennsylvania. He saw so many people relapse that cancer began to seem less like a biological problem and more like witchcraft. "It's like an evil entity that just keeps recurring and overcoming your best efforts." But when he began thinking about cancer from an evolutionary perspective, the problem became tractable again, he says.

Gatenby began trying to mathematically model the disease to work out how best to tackle it. His models suggest that many oncologists are taking the wrong approach. Typically, physicians will give the maximum dose of chemotherapy that a person can tolerate, to kill as many cancerous cells as possible. The hope is that they can wipe out the cancer before resistance evolves.

But studies from recent years suggest that tumours harbour drug-resistant cells long before they encounter therapy[5–7]. The population of resistant cells stays small because resistance comes with a fitness cost. When a patient receives a hefty dose of chemotherapy, however, the resistant cells become much fitter than the susceptible cells. Gatenby likens drug resistance to an umbrella: "If it's raining, the umbrella is very useful. But if it's not raining, it's a burden." Gatenby thinks that he can capitalize on the natural competition between susceptible and resistant cells by managing drug dosage or timing more carefully.

Recently, he tested the idea in mice with two kinds of breast cancer[8]. When he and his colleagues gave the mice the standard, maximum tolerated dose of the chemotherapy drug paclitaxel, the tumours roared back as soon as the treatment was stopped. The team also tried skipping doses whenever the tumour began to shrink, but that worked no better. A third group of mice received the standard high dose of chemotherapy at first, but once the animals' tumours started to shrink, the researchers dialled back the dose. This strategy resulted in the best survival for the mice and allowed three out of the five mice tested to be weaned off the

drug completely. The treatment is meant to adapt to how the tumour responds and maintain a balance between drug-resistant and susceptible cells (see 'Evolving strategies'). "I think it's one of the most exciting advances in cancer biology because it's a relatively easy thing to try," says Carlo Maley, a biologist at Arizona State University in Tempe who has collaborated with Gatenby.

In May 2015, the Moffitt Cancer Center launched a pilot study to test whether this kind of adaptive-therapy approach might help people with prostate cancer. Physicians will monitor the patient's levels of prostate specific antigen (PSA), a marker of disease progression. They will then administer standard treatment or stop it, depending on what they see. Researchers have studied intermittent therapy in the past, but the protocols generally involve rigidly controlled cycles. "With adaptive therapy, the on–off cycle is determined by the tumour response," Gatenby says. He also plans to use the wealth of molecular and clinical data from the trial to develop computer models that might guide adaptive therapy in the future.

IN A BIND

Physicians have noticed other evolutionary paradigms at work. In January, Jeffrey Engelman, a thoracic oncologist at Massachusetts General Hospital in Boston, and his colleagues detailed the case of a 52-year-old woman with metastatic lung cancer in *The New England Journal of Medicine*[9]. The woman's tumour had a genetic rearrangement that produced a misshapen version of the ALK protein, so her doctors first administered the drug crizotinib, which inhibits the action of ALK. She responded well for 18 months, but then relapsed. A second-generation therapy also failed, so physicians moved onto a third-generation therapy that is still in clinical trials. That worked for a while, but after less than a year, the woman's liver began to fail, and she had to be hospitalized. Then her doctors found that the third therapy had prompted a new mutation that made her cancer once again responsive to crizotinib. When they administered the drug, her liver recovered, and she improved so much that she was able to leave the hospital.

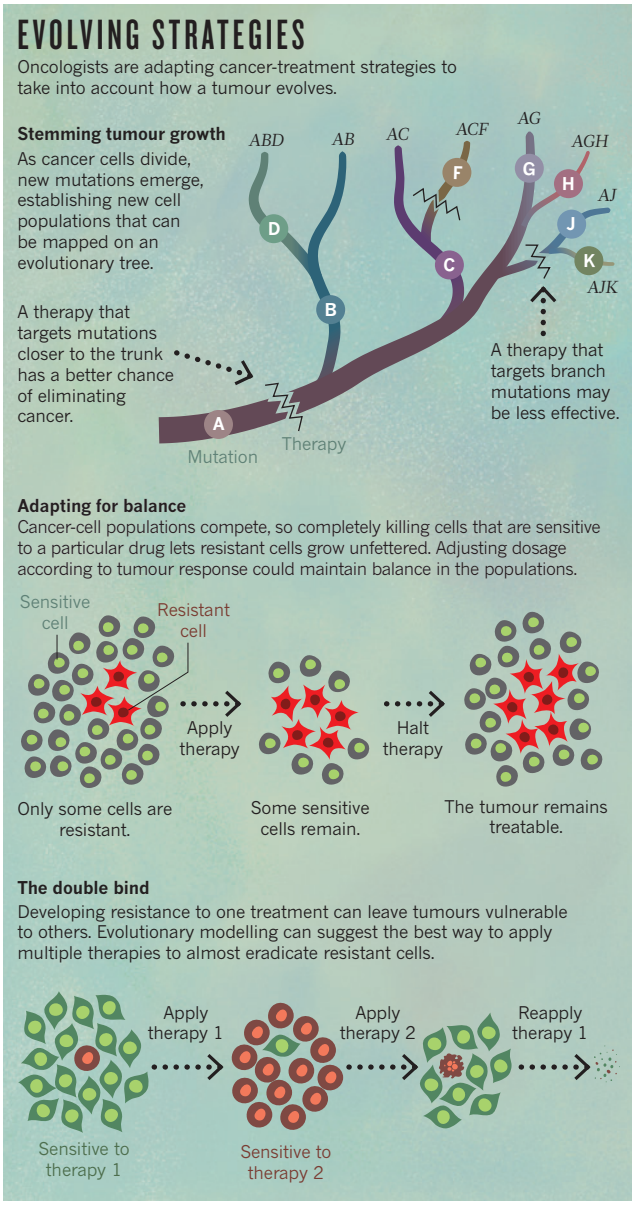
For Engelman and his colleagues, the woman's resensitization to crizotinib was a happy accident. But researchers may be able to direct cancer down such routes intentionally. Gatenby calls this strategy an evolutionary double bind, and he explains it like this: imagine trying to control a population of rats by introducing predators, such as hawks, that can pick them off from the sky. That type of predation selects for rodents that hide under brush. So one might bring in snakes that also hide under brush. The snakes would select for rats that prefer open spaces, making them vulnerable to hawks, Gatenby says. The same idea could apply in cancer. Use one treatment that makes the cancer more vulnerable to a second one, and then alternate between the two. It's "not whack-a-mole", Gatenby says, "but rather a carefully thought-out method that exploits the evolutionary dynamics".

That strategy is exactly what Ben Solomon, a cancer researcher at the Peter MacCallum Cancer Centre in Melbourne, Australia, plans to test in an upcoming trial. Many people with lung cancer harbour mutations in a gene called *EGFR*. Several drugs have been approved to target *EGFR* mutations, but tumours invariably develop resistance to them. In about half of patients, this resistance is caused by a mutation in *EGFR* called T790M. Last year, the US Food and Drug Administration approved a targeted drug called osimertinib, which inhibits the standard *EGFR* mutations as well as T790M, but people who respond to it tend to relapse within a year.

Solomon and his colleagues plan to start trial participants on osimertinib and then monitor resistance by tracking tumour DNA that circulates in their blood. The researchers expect to see a reduction in the T790M mutation. When that happens, they will switch to a first-generation EGFR inhibitor, which doesn't inhibit T790M. When T790M levels rise, the researchers will switch back to osimertinib. "Our hypothesis is that that's going to delay the emergence of resistance to osimertinib, because we're not maintaining that selection pressure," says Solomon. He hopes to have final approval for the trial soon.

There is no guarantee that any of these strategies will work. But even if the trials fail, the results of the tests will help researchers to refine their theories, and will address some of the big unknowns. How do the genetically diverse cells in a tumour interact, for example, and what is the role of the cellular environment that they inhabit? Kornelia Polyak, an oncologist at Harvard Medical School in Boston, says that cancer researchers tend to focus on the mutations inside cells, and fail to consider how those mutated cells might influence the cells around them. "That's a largely unexplored area," she says.

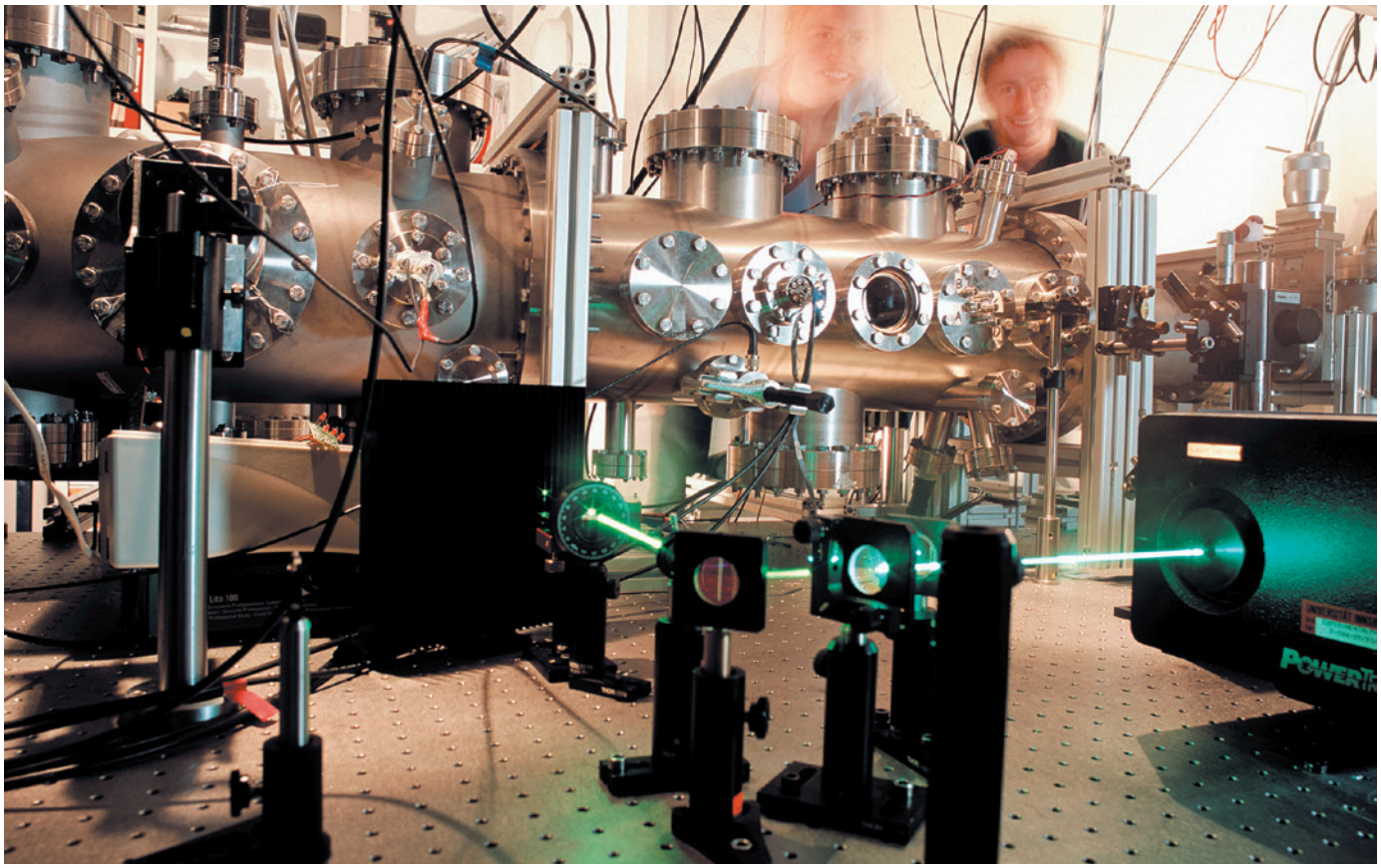
The dynamics inside a tumour are exceedingly complicated, but Engelman is not discouraged. Clinical analyses will help researchers to understand this complexity. "These insights are going to bring us closer to having bigger and bigger impacts," he says. "The depressing thing is to not know what the hell is going on." ■



Cassandra Willyard *is a science writer based in Madison, Wisconsin.*

1. Misale, S. *et al. Nature Commun.* **6,** 8305 (2015).
2. Gerlinger, M. *et al. N. Engl. J. Med.* **366,** 883–892 (2012).
3. Bozic, I. *et al. eLife* **2,** e00747 (2013).
4. McGranahan, N. *et al. Science* **351,** 1463–1469 (2016).
5. Turke, A. B. *et al. Cancer Cell* **17,** 77–88 (2010).
6. Bhang, H. C. *et al. Nature Med.* **21,** 440–448 (2015).
7. Su, K. Y. *et al. J. Clin. Oncol.* **30,** 433–440 (2012).
8. Enriquez-Navas, P. M. *et al. Sci. Transl. Med.* **8,** 327ra24 (2016).
9. Shaw, A. T. *et al. N. Engl. J. Med.* **374,** 54–61 (2016).

COMMENT



VOLKER STEGER/SPL

A quantum-entanglement experiment at the University of Vienna in Austria.

Unite to build a quantum internet

Advances in quantum communication will come from investment in hybrid technologies, explain **Stefano Pirandola** and **Samuel L. Braunstein**.

Almost 25 years ago, physicists discovered a way of 'teleporting' a quantum system from one place to another without moving it[1]. There are physical limits to such teleportation: nothing can be transmitted faster than the speed of light; and Heisenberg's uncertainty principle restricts what we can know about the state of a quantum system at a given time. Nevertheless, the transported system is a replica that perfectly mimics the original, thanks to the weirdest feature of quantum mechanics — entanglement. Described by Albert Einstein as "spooky action at a distance", this property enables distinct quantum systems to become intimately correlated so that an action performed on one has an effect on the other, even for systems that are too far apart to physically interact.

Quantum states are fragile and cannot be sent through conventional lines of communication; quantum teleportation offers a reliable and efficient way to transfer quantum information across a network. It provides the most promising mechanism for a future quantum internet, with ▶

secure communications and a distributed computational power that greatly exceeds that of the classical Internet.

Quantum information comes in a variety of forms — the polarization state of a photon, the spin of an electron or the excitation state of an atom. Many technologies have been developed for teleporting such states[2]. But there are practical restrictions on what can be teleported, and how. Certain technologies will be better than others for particular tasks, and each has its limitations. Polarized photons have been used to transfer quantum information over more than 100 kilometres[3], but only probabilistically. Superconducting devices can send information without losses through a chip, but only for a split second, after which the information is scrambled by interactions with the environment.

Hybrid approaches might overcome these limitations. A global, distributed quantum computer or internet will need to integrate different sorts of quantum technologies. For example, light-based teleportation, for long-distance quantum communication, will need to be linked to matter-based quantum memories and quantum computers for data storage and data processing. Here, we outline the main challenges and call for researchers to focus on the interfaces between quantum technologies as well as advancing individual methods.

TWO APPROACHES

The best technique at present for long-distance communication is the teleportation of quantum information that is embedded in optical light. Quantum information — measured in units known as quantum bits, or qubits — can be encoded either by the discrete properties of a pulse of light, such as its polarization state, or by the continuous aspects of an electromagnetic wave, such as the intensity and phase of the wave's electric field[4]. To teleport this information, both the sender and receiver must own one of a pair of entangled quantum systems (see 'Quantum teleportation'). When the sender alters the state of their system, the receiver's system is also affected.

Polarization qubits perform best in terms of distance — holding a record of 143 kilometres[3]. But currently, only 50% of these qubits can be teleported[2]. Teleportation requires that the sender can carry out an operation known as a Bell detection, in which the polarizations of two qubits are correlated perfectly in four possible configurations. But there is no practical way to measure all four outcomes. Simple optics and photodetectors can distinguish two, at most. Extra qubits add technical complications[5].

Inconclusive outcomes such as these are acceptable for quantum cryptography, in which secret keys are generated at random, and part of the information can be discarded. But quantum communication demands that information is sent in full.

Teleportation over long distances[2] brings further technical challenges, such as compensating for atmospheric turbulence and movement of the ground. It is also likely to require new technologies to synchronize both ends — using atomic clocks, for example. Modern classical communications rely heavily on satellite technology. Transferring quantum information to the ground from a satellite in a low Earth orbit (at an altitude of about 500 kilometres) is within the reach of current technology, thanks to ground-based telescopes with metre-sized apertures that can collect most of the light from a beam that has spread out during its passage from a satellite. But transferring quantum information from the ground to a satellite, or between satellites, is more difficult because satellites cannot carry large optics.

By contrast, it is easy to measure all Bell-detection outcomes for continuous-variable systems such as electric fields, using only simple linear optics and standard photodetectors. Such systems can convey simultaneously the equivalent of many qubits, which makes them appealing for use in high-rate quantum communications[6]. But because the range of distances over which they can teleport is currently limited, continuous-variable systems are used less frequently than are qubits.

Approaches are needed that combine the best features of discrete variables (teleportation over long distances) with those of continuous variables (fast, deterministic teleportation). Teleportation that uses such combinations has been demonstrated over table-top distances. One experiment[7] combined a discrete qubit with a continuous-variable entangled source to teleport quantum information deterministically. Further studies should help both to extend the distances covered by these experiments and to integrate qubits with other types of quantum technologies, including quantum memories for the storage of teleported information.

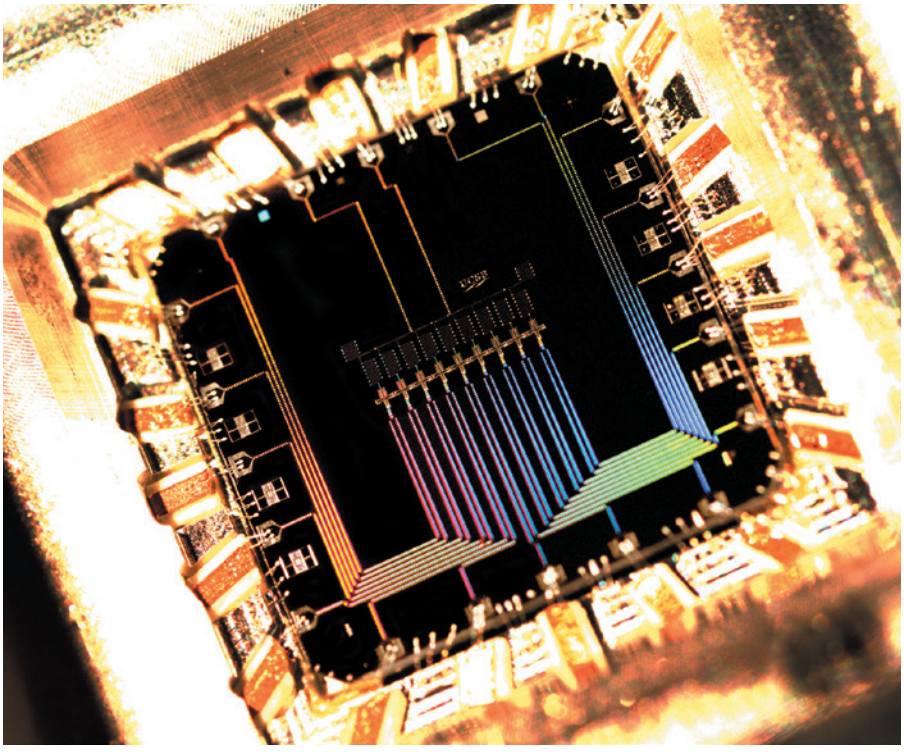
Studies of hybrid technologies will require greater collaboration and interaction between teams with different specializations.

QUANTUM INTERNET

One of the greatest challenges for implementing a globally distributed quantum computer or a quantum internet is entangling nodes across the network[8]. Qubits can then be teleported between any pair and processed by local quantum computers.

Ideally, nodes should be entangled either in pairs or by creating a large, multi-entangled 'cluster state' that is broadcast to all nodes. Cluster states that link thousands of nodes have already been created in the laboratory[9]. The challenges are to demonstrate how they might be deployed over long distances, as well as how to store quantum states at the nodes and update them constantly using quantum codes.

Quantum networks require memories to store quantum information, ideally for hours — shielding it from unwanted interactions with the environment. Such memories are

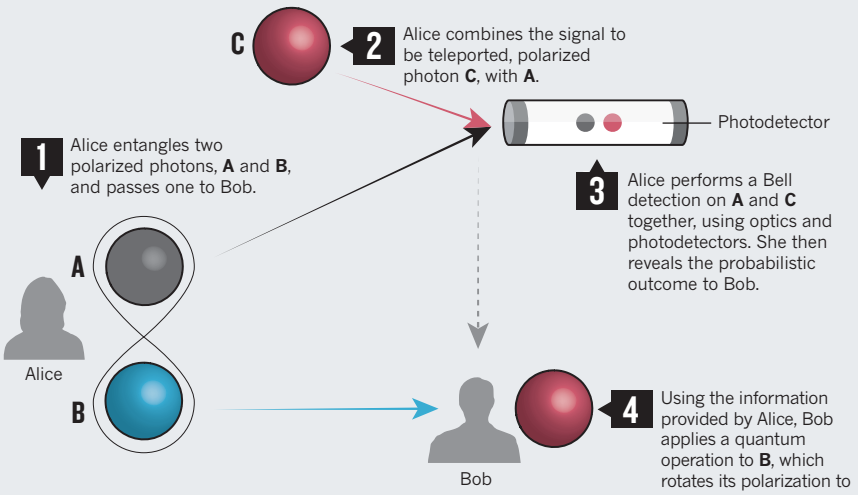


JULIAN KELLY/MARTINIS GROUP

A superconducting quantum chip with nine qubits.

QUANTUM TELEPORTATION

A signal, such as the polarization state of a photon, can be teleported from one place to another using entangled photons and quantum measurements. (Physicists often call the sender Alice and the receiver Bob.)



needed for quantum computing at nodes and also for the faithful, long-distance distribution of entanglement through quantum repeaters.

Quantum memories need to convert electromagnetic radiation into physical changes in matter with near-perfect read–write fidelity and at high capacity. 'Spin ensembles' represent one type of quantum memory. Ultracold atomic gases consisting of about one million atoms of rubidium can convert a single photon into a collective atomic excitation known as a spin wave. Storage times are approaching the 100 milliseconds required to transmit an optical signal across the world.

Solid-state quantum memories are even more appealing. Crystalline-solid spin ensembles — created by inserting lattice defects known as nitrogen-vacancy centres into diamonds, or by doping rare-earth crystals — can remain coherent for hours at cryogenic temperatures.

Superconducting qubits, which are defined by physical quantities such as the charge of a capacitor or the flux of an inductor, interact within a quantum processor by releasing and absorbing microwave photons. For the successful integration of solid-state quantum memory, reversible storage and retrieval of quantum information must be made possible. This will require an efficient interface between the microwave photons and the atomic spins of a solid-state quantum memory that is attached to the processor. If successful, this hybrid technology would become the most promising architecture to be scaled up into a large, distributed quantum computer.

The incorporation of superconducting processors into a quantum internet also requires that locally processed and stored microwave photons interface with optical signals (often carried in fibres), which are the most robust carriers of quantum information over long distances. A hybrid solution, known as an optomechanical quantum transducer, is emerging[10]. These devices exploit nanomechanical oscillators (such as microscopic vibrating mirrors) to transform optical photons into microwave photons, and vice versa. But their efficiency must be improved to ensure that qubits are not lost during the conversion process and that all of their quantum features are preserved. The conversion efficiency is currently about 10% (ref. 10).

"The development of a quantum internet needs investment on a much larger scale."

The next 15 years could see the construction of a hybrid-technology quantum internet. In the vision we outline, superconducting quantum processors will be integrated with solid-state memories for local quantum storage and then augmented with microwave–optical transducers for long-distance optical communication. After two remote nodes have been connected in this way, entanglement can be distributed between distant quantum processors to enable teleportation.

NEXT STEPS

To make this vision a reality, the following three steps should be priorities for quantum teleportation science.

First, more research — theoretical and experimental — is needed at the interface between discrete and continuous variables; dedicated conferences would help. This would enable us to blend these currently distinct approaches to exploit the best of both. Satellite experiments with polarization qubits should be pursued, and continuous-variable teleportation should be extended beyond the lab for communication within cities using free space or optical fibres.

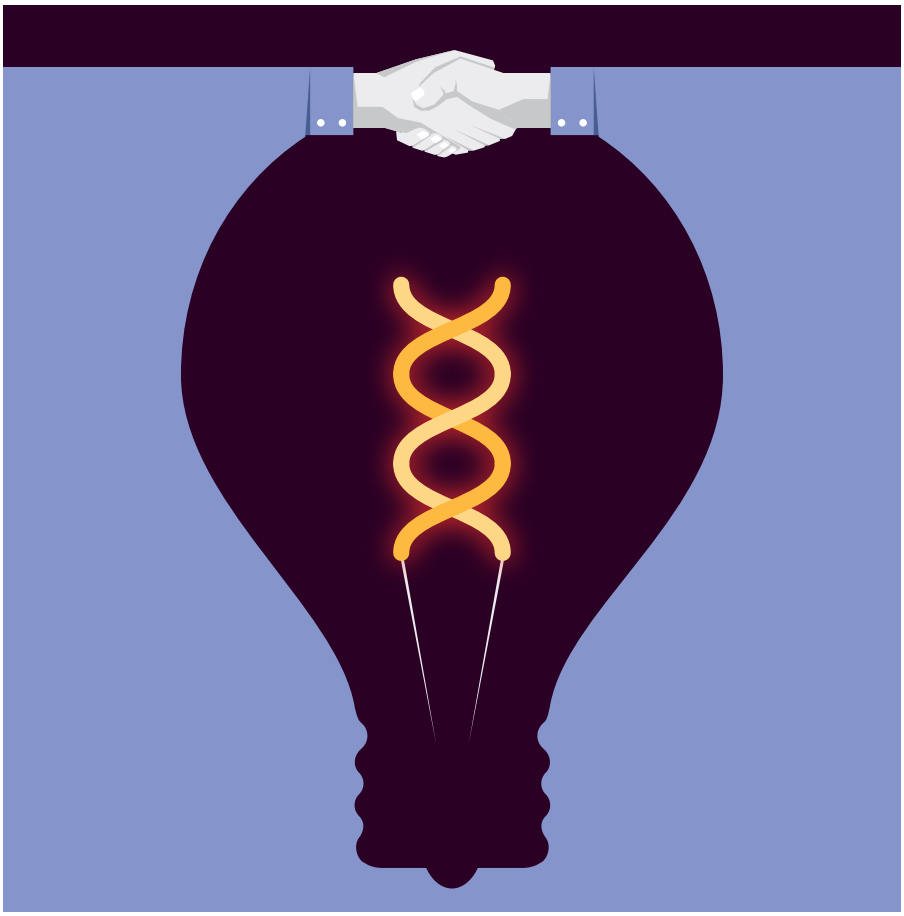
Second, the most successful technologies will be those that integrate data communication and data storage. We need to invest in the development of a more efficient interface between superconducting quantum processors and solid-state quantum memories. This would improve the performance of the storage and retrieval of microwave photons. A tangible next step could be on-chip teleportation between a superconducting qubit and a nitrogen-vacancy centre in a local quantum memory.

Third, investment should be made in technologies that show promise of scalability. For example, microwave–optical transducers that can efficiently connect microwave photons with optical photons on a chip for long-distance quantum communication should be designed and integrated. Two remote chips could be linked by paired transducers, paving the way for long-distance quantum teleportation between superconducting qubits.

These steps will necessitate a closer interaction between researchers in superconducting quantum computing and those who are developing long-distance quantum optical communications. Industry must also be involved, especially multinational corporations that are leaders in computer hardware and telecommunications. Quantum technology is attracting private stakeholders, but the development of a quantum internet needs investment on a much larger scale. ■

Stefano Pirandola *is reader and* **Samuel L. Braunstein** *is professor in the Department of Computer Science at the University of York, UK.*
e-mails: stefano.pirandola@york.ac.uk; sam.braunstein@york.ac.uk

1. Bennett, C. H. *et al. Phys. Rev. Lett.* **70,** 1895–1899 (1993).
2. Pirandola, S., Eisert, J., Weedbrook, C., Furusawa, A. & Braunstein, S. L. *Nature Photon.* **9,** 641–652 (2015).
3. Ma, X.-S. *et al. Nature* **489,** 269–273 (2012).
4. Weedbrook, C. *et al. Rev. Mod. Phys.* **84,** 621–669 (2012).
5. Knill, E., Laflamme, R. & Milburn, G. J. *Nature* **409,** 46–52 (2001).
6. Pirandola, S. *et al. Nature Photon.* **9,** 397–402 (2015).
7. Takeda, S., Mizuta, T., Fuwa, M., van Loock, P. & Furusawa, A. *Nature* **500,** 315–318 (2013).
8. Kimble, H. J. *Nature* **453,** 1023–1030 (2008).
9. Yokoyama, S. *et al. Nature Photon.* **7,** 982–986 (2013).
10. Andrews, R. W. *et al. Nature Phys.* **10,** 321–326 (2014).



ADAPTED FROM HONG LI/GETTY

Pursuit of profit poisons collaboration

The CRISPR–Cas9 patent battle demonstrates how overzealous efforts to commercialize technology can damage science, writes **Jacob S. Sherkow**.

Last month, in an extraordinary dispute before the US Patent and Trademark Office (USPTO), university lawyers laid out their clients' legal strategies for claiming patents that cover the celebrated gene-editing technology CRISPR–Cas9. Over the next year, the USPTO will receive volumes of evidence centred on who first invented the technology.

Battles over scientific priority are as old as science itself. But the CRISPR–Cas9 patent dispute is unusual because it pits two leading research institutions against one another for the control and industrial development of a foundational technology: the University of California, Berkeley (UC Berkeley), and the Broad Institute of MIT and Harvard in Cambridge, Massachusetts.

As scientific institutions increase their involvement in the commercialization of research[1], it is worth considering the potential consequences for science if more institutions follow the path of UC Berkeley and the Broad Institute.

HIGH STAKES

In May 2012, researchers at UC Berkeley, led by Jennifer Doudna and her collaborator, Emmanuelle Charpentier (then located at the University of Vienna in Austria) filed a patent application in the United States for CRISPR–Cas9. Seven months later, Feng Zhang, a researcher at the Broad Institute, filed a competing application that covered similar uses of the technology. After Zhang's lawyers requested that his application be fast-tracked, the USPTO awarded one patent to Zhang in April 2014, followed by a dozen more in the subsequent 12 months. Meanwhile, the application made by Doudna and her colleagues languished.

Last April, Doudna's lawyers requested that the USPTO conduct a specialized legal trial, known as a patent interference, to determine the ownership of the US patents that cover the CRISPR–Cas9 system. This January, the USPTO formally agreed to carry out the proceeding.

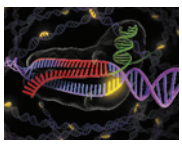
One conspicuous aspect of this case, in my opinion, is the degree to which UC Berkeley and the Broad Institute have weighed in on what is essentially a dispute over scientific priority.

The Broad Institute has produced press releases, videos and a slick feature on its website that stress the importance of Zhang's contributions to the development of the CRISPR–Cas9 technology. And earlier this year, the central positioning of Zhang's work in a historical perspective of CRISPR published in *Cell*[2] by the president and director of the Broad Institute, Eric Lander, prompted a storm of angry responses from scientists, including Doudna and Charpentier. Meanwhile, at UC Berkeley, a press release that discussed the potential of CRISPR described Doudna as "the inventor of the CRISPR–Cas9 technology" (see go.nature.com/cm2gvx).

The financial stakes are high. The CRISPR–Cas9 patents are widely viewed to be worth hundreds of millions, if not billions, of dollars. Both organizations have invested directly in spin-off companies that were co-founded by their researchers — the Broad Institute in Editas Medicine, co-founded by Zhang, and UC Berkeley in Caribou Biosciences, co-founded by Doudna. A report submitted by Editas in January to the US Securities and Exchange Commission lists the Broad Institute and other Harvard-affiliated institutions as owning a major equity stake in the company: about 4.2% of its common shares (see go.nature.com/45c1ey).

DIFFERENT TIMES

Efforts to commercialize the research output from universities played out differently in the past. Since 1980, US universities have been able to patent the inventions of their researchers, thanks to the Bayh–Dole Act — legislation that determines the ownership of intellectual property arising from federally funded research. But for the most part, institutions have kept their distance from disputes over scientific priority. In fact, after factoring in the costs of filing patents and staffing, university technology-transfer offices have generally been money losers for their institutions[3].



NATURE.COM
For more of *Nature*'s coverage on CRISPR, see:
nature.com/crispr

L: RICK FRIEDMAN/CORBIS; R: JOHN ELK III/LONELY PLANET/GETTY

Even in the case of lucrative patents, commercial development has frequently been left to venture capitalists and the researchers themselves. Take the Cohen–Boyer patents, which covered early gene-splicing technology and netted Stanford University and the University of California, San Francisco (UCSF), both in California, hundreds of millions of dollars in licensing fees during the 1980s and 1990s. In this instance, Genentech, the company in South San Francisco, California, that was formed to commercialize the underlying technology, sprung from the efforts of Herbert Boyer, one of the founding researchers, and the financier Robert Swanson. The company was neither owned by, nor an exclusive licensee of, Stanford or UCSF.

Research institutions in general are starting to play a bigger part in shepherding their researchers' projects through the commercialization process. A 2014 report from the Association of University Technology Managers in Oakbrook Terrace, Illinois — an organization that supports managers of intellectual property at academic research institutions, non-profit organizations and government agencies worldwide — documented that universities are increasing equity investments in their researchers' start-up companies. Of the patent licences granted by universities in 2014, 10% were tied to such investments[1], compared with 6.7% in 1999 (ref. 4).

I am concerned that such involvement in commercialization has the potential to clash with the broader, educational mission of research institutions.

Universities worldwide have long strived to foster a culture of scientific collaboration. Even when universities have obtained broad patents, as the Carnegie Institute of Washington in Washington DC did in the early 2000s for a gene-expression control technology known as RNA interference, licences have been cheap and easy for researchers to obtain[5]. In other cases, scientists have simply ignored patents that cover fundamental technologies[6].

Academic research institutions now seem less shy about taking each other to court for patent infringement. In 2011, the University of Utah in Salt Lake City sued the Max Planck Society for the Advancement of Science in Germany over claims to a patent that covered a technology called short interfering RNA, which inhibits gene expression (see go.nature.com/vyujnp). And over the past four years, Stanford University and the Chinese University of Hong Kong in Sha Tin have engaged in a heated patent litigation over prenatal genetic diagnostic blood tests, a market that was worth US$530 million in 2013.

In the current era of budget tightening, universities of all stripes might be tempted to use licensing fees as another funding mechanism. The University of South Florida in Tampa, for example — a public institution that had its state funding cut by $48 million in 2012 — holds a substantial number of patents that have not yet been licensed and has a famously low ratio of patent-licence revenue to research expenditure[7]. If its financial situation were to deteriorate further, the university might be compelled to extract licence fees from other research institutions for those patents.

The Broad Institute of MIT and Harvard (left) and the University of California, Berkeley (right).

PATH TO PROFIT

It would be wrong to suggest that patents, writ large, are failing educational research institutions. In the cases of gene splicing, RNA interference and human embryonic stem cells, patents have been major earners for institutions and researchers without damaging the scientific enterprise[5].

"Efforts to commercialize the research output from universities played out differently in the past."

But an obvious danger of increasing the focus on commercialization is that educational institutions will view scientific research as a path to profit, above all else. It is not hard to imagine that patent disputes might lead to university administrators pushing certain views on their scientists, denigrating collaboration with researchers from competing institutions and tasking tenure committees with valuing patents over publications.

Where scientific advances have the potential to be profitable, universities should support researchers to bring that work to fruition. This might include helping them to secure patents. But it is my view that serious commercialization efforts — such as granting exclusive licences or receiving equity ownership in researchers' start-ups — should be left to industry.

The CRISPR–Cas9 dispute could have played out very differently. Zhang and Doudna were both co-founders of Editas. And UC Berkeley and the Broad Institute could have filed patent applications that listed the research teams from both institutions as co-inventors. Any resulting patents could then have been freely or cheaply licensed to other research institutions, or used to fund a joint academic organization dedicated to studying the technology. The patents could also have been widely, but not exclusively, licensed to a variety of industry competitors — promoting a robust, competitive market for commercial CRISPR–Cas9 applications and creating a funding stream for further academic research.

Biomedical research in educational institutions has long prided itself on a culture of openness and sharing — one that both Zhang and Doudna have exercised by donating various components of the CRISPR–Cas9 system to the open-science consortium Addgene in Cambridge, Massachusetts. The incentives that patents create for educational institutions should not be allowed to erode scientific collaboration. ■

Jacob S. Sherkow *is associate professor of law, Innovation Center for Law and Technology, New York Law School, New York, USA.*
e-mail: jacob.sherkow@nyls.edu

1. Association of University Technology Managers. *AUTM US Licensing Activity Survey: FY2014* (AUTM, 2014).
2. Lander, E. S. *Cell* **164,** 18–28 (2016).
3. Abrams, I., Leung, G. & Stevens, A. J. *Res. Mgmt. Rev.* **17,** 18–50 (2009).
4. Association of University Technology Managers. *AUTM US Licensing Survey: FY1999* (AUTM, 1999).
5. Sherkow, J. S. *Nature Biotechnol.* **33,** 256–257 (2015).
6. Eisenberg, R. S. *Houston Law Rev.* **45,** 1059–1099 (2008).
7. Feldman, R. & Price, N. W. II *Stanford Tech. Law Rev.* **17,** 773–808 (2014).

Some birds evolve signature egg colours and patterns to confuse nest parasites such as the cuckoo finch. Each column shows eggs from one host species.

ELEANOR CAVES AND CLAIRE SPOTTISWOODE

ORNITHOLOGY

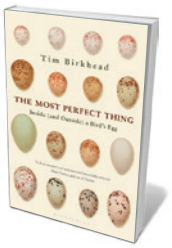
Oology unshelled

John M. Marzluff extols a rich history of ornithology's debt to egg collecting.

Tim Birkhead has spent much of the past four decades watching birds, and in particular mucking around guano-covered ledges on which seabirds breed. His insights have revolutionized ornithologists' understanding of mate fidelity; his ability to distil complex science for the general reader, for example in *Bird Sense* (Bloomsbury, 2013), has revealed what it is like to be a bird. Now, with an eye on past discoveries and persistent puzzles, *The Most Perfect Thing* reveals what it is like to become a bird — from nascent ovum to shelled egg and beyond.

Birkhead starts his story in the nineteenth century, on English cliffs where eccentric collectors wait anxiously for daring "climmers" to fetch unusually shaped and patterned eggs from the nests of common guillemots (*Uria aalge*). Especially prized was a sequence of eggs laid by the same bird throughout her life: each was identically marked with a design of splotches and scribbles. The extensive collections of early oologists, such as George Lupton, who amassed more than 1,000 guillemot eggs, fascinate Birkhead even as he laments this now illegal and inadmissible practice. He tracked down many old collections to learn more about the evolution of shape and colour. Through careful evaluation of alternative hypotheses, he dispels the common explanation that the pear shape of the guillemot's egg evolved mainly to keep it from rolling off the nest precipice. Rather, the shape probably provides legroom for the developing chick, enables the egg to tip above the faecal stew that often surrounds it, and increases surface area — improving heat transfer during incubation.

NATURE.COM
For more on science in culture see:
nature.com/booksandarts



The Most Perfect Thing: Inside (and Outside) a Bird's Egg
TIM BIRKHEAD
Bloomsbury: 2016.

With equal thoroughness, Birkhead shows that the unusual markings — from light peppering, squiggling and blotching to completely blackened ends or dark rings around the midline — that entranced collectors enable guillemot parents to recognize their own eggs. This, however, is only part of the story of egg colour. Other species' eggs are marked to camouflage them: the Japanese quail (*Coturnix japonica*), for example, lays heavily mottled eggs in nest sites with matching patterns. Some eggs, such as that of the ostrich (*Struthio camelus*), are white to protect the developing chick from the heat of the Sun. Others are brightly coloured, as with the American robin (*Turdus migratorius*), whose blue eggs advertise the quality of the brooding female. Still others may be lightly pigmented to raise their internal temperatures or to increase light penetration, which can speed up chick development. This mechanism may be used to synchronize hatching within a clutch: the last eggs laid are often the lightest in colour. Through the eye of this careful evolutionary ecologist, and a series of high-quality colour plates, we come to appreciate the beauty and functionality of eggs.

Having considered the whole egg, Birkhead next describes its making. He writes clearly, with accuracy and wit, about the ovum's development in the bird's ovary and its journey through the oviduct. We learn about the microstructure of the shell — much like a rigid sieve — and how pores and cuticle adapt

the egg to local atmospheric conditions, as well as repelling water, hydrophilic microbes and contaminants that can diffuse in and challenge the developing embryo. Foreign bodies that make it past the shell are dealt with in the albumen, which Birkhead describes as a "sophisticated biochemical firewall against microbes". He then explains how the yolk provides fats and proteins manufactured in the mother's liver to the growing chick; it also furnishes the chick with crucial antioxidants, vitamins and hormones such as testosterone.

The embryo's survival may be enhanced by the odorous, oily secretions of the mother's preen gland that grease the eggs' shells during laying and incubation. In some species, such as the hoopoe (*Upupa epops*), this oil mixes with the copious droppings in the nest to produce a notorious 'filth', which may include beneficial bacteria that enhance hatchability. Conservation biologists working to restore rare species may find that the lack of such bacteria helps to explain why many eggs are difficult to hatch without at least a modicum of parental incubation.

Along the way, Birkhead introduces many colourful characters little known to science. Sequences of eggs collected from the 1970s to the early 2000s by John Colebrook-Robjent in Zambia were crucial to our understanding of the 'arms race' in egg coloration between nest parasites and their hosts. Birds such as cuckoo finches (*Anomalospiza imberbis*), for instance, sneak their eggs into the nests of others, but must continually adjust the background colour and pattern to fool the hosts — which in turn evolve unique markings to foil the parasite's disguise. Lupton's guillemot-egg collection inspired and informed current understanding of the adaptive value of egg shape. British physician Allen Thomson speculated that the yolk is built up in layers. Birkhead's ability to weave together history and science shows the human nature of research.

He has a marvellous way with words, writing of monogamous albatrosses living like "long-distance truck drivers — at home with their partner only occasionally and making the most of it when they are". And he tantalizes with unsolved mysteries. Why, for example, does the egg of a chicken travel through the hen pointed end first until the very last minute, when it turns through 180° on the horizontal plane to be laid blunt end first?

Birkhead's historical acumen and sharp pen had me seeing eggs in a new light. He has convinced me that they are splendid, if not indeed most perfect. ■

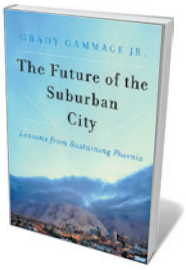
John M. Marzluff *is the James W. Ridgeway Professor of Wildlife Science in the School of Environmental and Forest Sciences at the University of Washington, Seattle. His latest book is* Welcome to Subirdia.
e-mail: corvid@uw.edu

Books in brief

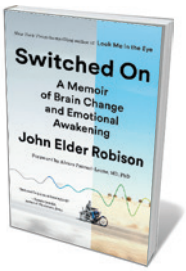
The Great Departure: Mass Migration from Eastern Europe and the Making of the Free World
Tara Zahra W. W. Norton *(2016)*
As many as 58 million Europeans seeking "bread and freedom" poured into the Americas from 1846 to 1940. Millions returned — worn down by the punishing, ill-paid labour driving the New World's booms. Historian Tara Zahra's timely, myth-busting chronicle shows how, early on, European states attempted to "scientifically" manage masses of people to serve their own and international goals. The impacts ranged all the way from the Holocaust to a shift in the concept of freedom, to the right to stay or leave.

15 Million Degrees: A Journey to the Centre of the Sun
Lucie Green Viking *(2016)*
Earth may be 150 million kilometres from the Sun, but few relationships are as intimate. Aside from the star's centrality to life, our planet is embraced by the magnetic bubble of the heliosphere. Solar physicist Lucie Green's engrossing primer clearly explicates the science and its star-studded history. That stretches from Galileo's work on sunspots to astronomer Cecilia Payne-Gaposchkin's 1925 discovery that helium is the most abundant element in the Sun — and physicist Sami Solanki's 2004 finding that the past 70 years of grand-maximum solar activity may be a rare blip.

Show Me the Bone: Reconstructing Prehistoric Monsters in Nineteenth-Century Britain and America
Gowan Dawson University of Chicago Press *(2016)*
A putative knack for mentally constructing a beast entire from a scattering of fossilized remains lent early palaeontologists a sorcerer-like glamour. As the field's founder, Georges Cuvier, put it: "Give me the bone, and I will show you the animal." Science historian Gowan Dawson lucidly traces the afterlife of Cuvier's incorrect "law of correlation" in Victorian Britain and the United States. The idea seeped into science, irking biologist T. H. Huxley and, argues Dawson, subtly influencing Charles Darwin's thinking on natural variation.



The Future of the Suburban City: Lessons from Sustaining Phoenix
Grady Gammage Jr Island *(2016)*
The car, the shopping centre and the single-family home, reveals urban specialist Grady Gammage Jr, created the "suburban city" — sprawls clustered in the US southwest and associated with rampant development. In his study of their potential for sustainable transition, Gammage focuses on Phoenix, Arizona — vast, traffic-ridden and caught between aridity and a per capita water consumption more than twice that of New York. He argues that its historic reliance on renewable surface water rather than groundwater, and openness to low-carbon light transport, point to a potential for future resilience.



Switched On: A Memoir of Brain Change and Emotional Awakening
John Elder Robison Spiegel & Grau *(2016)*
In 2007, John Elder Robison published *Look Me in the Eye* (Crown), a raw memoir about growing up with Asperger's syndrome. The following year, cognitive neurologist Alvaro Pascual-Leone invited him to participate in a study involving transcranial magnetic stimulation (TMS). Here, Robison chronicles the "powerful mojo" that ensued as his emotions, empathy and perceptions deepened, colouring work and intimate relationships unexpectedly, even after the TMS effects faded. The science and ethical quandaries are deftly interlaced. **Barbara Kiser**

EVOLUTION

Taxonomies of cognition

Joan B. Silk examines Frans de Waal's treatise on the evolution of animal intelligence.

In *Are We Smart Enough to Know How Smart Animals Are?*, ethologist Frans de Waal celebrates the evolution of intelligence in nature. His is an entertaining account of how octopuses escape from jars by unscrewing the lids and rooks drop pebbles into a tube to access floating rewards. Natural selection, he argues, shapes cognitive abilities in the same way as it shapes traits such as wing length. As animals' challenges and habitats differ, so do their cognitive abilities. This idea, which he calls evolutionary cognition, has gained traction in psychology and biology in the past few decades.

For de Waal, evolutionary cognition has two key consequences. First, it is inconsistent with the concept of a 'great chain of being' in which organisms can be ordered from primitive to advanced, simple to complex, stupid to smart. Name a 'unique' human trait, and biologists will find another organism with a similar one. Humans make and use tools; so do wild New Caledonian crows (*Corvus moneduloides*). Humans develop cultures; so do humpback whales (*Megaptera novaeangliae*), which socially transmit foraging techniques. We can mentally 'time travel', remembering past events and planning for the future; so can western scrub jays (*Aphelocoma californica*), which can recall what they had for breakfast on one day, anticipate whether they will be given breakfast the next and selectively cache food when breakfast won't be delivered.

Furthermore, humans do not necessarily outdo other animals in all cognitive domains. Black-capped chickadees (*Poecile atricapillus*) store seeds in hundreds of locations each day, and can remember what they stored and where, as well as whether items in each location have been eaten, or stolen. Natural selection has favoured those prodigious feats of memory because they spell the difference between surviving winter and starving before spring. Human memory doesn't need to be as good: primates evolved in the tropics. "In the utilitarian view of biology," de Waal argues, "animals have the brains they need — nothing more, nothing less."

The second consequence of de Waal's view is that there is continuity across taxa. One source of continuity is based on evolutionary history: natural selection modifies traits to create new ones, producing commonalities among species with a common history. He points out that tool use is found not just in humans and chimpanzees, but also in other apes and monkeys, implying that relevant cognitive building blocks are shared across all primates. Continuity is also generated by convergent evolution, which produces similar traits in distantly related organisms such as New Caledonian crows and capuchin monkeys. De Waal opines that continuity "ought to be the default position for at least all mammals, and perhaps also birds and other vertebrates".

Are We Smart Enough to Know How Smart Animals are?
FRANS DE WAAL
W. W. Norton: 2016.

He calls for a moratorium on claims of human uniqueness, arguing that their proponents have overvalued human complexity, or undervalued that of other species. And he is correct that such claims have been repeatedly refuted — and often have a nonscientific basis. Charles Darwin's *On The Origin of Species* (1859) and *The Descent of Man* (1871) may be 150 years behind us, but many people remain uncomfortable with the view that humans are the product of the same processes that shaped other organisms. As the Bishop of Worcester's wife reportedly exclaimed when she heard of Darwin's theory: "Dear me, let us hope it is not true. But if it is true, let us hope it does not become widely known." And some who acknowledge natural selection's role in our origins are less comfortable with the idea that it has important effects on how we think, feel and behave. Efforts to introduce evolutionary perspectives into anthropology and psychology in the 1980s met fierce resistance and remain controversial.

But anthropocentrism, or what de Waal calls "anthropodenial", is not the only reason researchers are eager to understand what is distinctly human; some are driven by curiosity about how humans came to dominate the planet. The biomass of humans and domesticated animals exceeds the biomass of all wild vertebrate species. Our success presumably has something to do with the emergence of a unique suite of cognitive traits.

A woodpecker finch (*Camarhynchus pallidus*) uses a stick for foraging.

OLE JORGEN LIODDEN/NATURE PICTURE LIBRARY

De Waal recognizes only one such trait: our rich and flexible system of symbolic communication, and our ability to exchange information about past and future. His commitment to the principle of continuity forces him to discount the importance of language for human cognition because of evidence of thinking by non-linguistic creatures. And he ignores compelling findings from linguists and developmental psychologists such as Elizabeth Spelke on the formative role of language in cognition.

De Waal pays little attention to the evolutionary processes that create inter-species differences. Every species is a mixture of traits inherited from ancestral taxa and derived traits that evolved after the species diverged onto its own path. So colour perception is due to visual pigments made of opsins, proteins sensitive to particular wavelengths of light. Most mammals have only two and cannot distinguish between red and green. *Homo sapiens* can because of the duplication and modification of an opsin gene in the common ancestor of apes and Old World monkeys, which all have three such genes. Derived traits produce real discontinuities between species.

A better book would have celebrated both similarities in the foundations of cognitive abilities across species, and processes that produce differences in cognitive abilities between species. A more useful book would have included some discussion of mechanisms (such as perception) that underlie cognitive abilities in different taxa. A more balanced book would not categorize as "killjoy accounts" all sceptical accounts of de Waal's favoured cognitively generous interpretations of behaviour, or summarily reject negative evidence from well-designed experiments. A more satisfying book would leave readers with a clearer understanding of why, a few million years after our lineage diverged from the lineage of chimpanzees, we are the ones reading this book, and not them. ■

Joan B. Silk *is in the Institute of Human Origins at Arizona State University, Tempe.*
e-mail: joansilk@gmail.com

CORRECTION
In the essay 'Getting the circulation going' (*Nature* **531,** 443–446; 2016), William McDonough was said to have studied under John Lyle; in fact, they collaborated.

Correspondence

Climate costing is politics not science

Nicholas Stern argues that today's integrated assessment models for quantifying the economic and societal impacts of climate change are inadequate (*Nature* **530,** 407–409; 2016). We disagree with his view on the superiority of more complex models such as DSGE (dynamic stochastic computable general equilibrium) models, which purport to account for a larger class of uncertain future events.

In our view, DSGE models have proved to be ineffective for policymaking, even in simple, short-term settings of pure economics, by failing to anticipate the onset of the recent recession (see P. Mirowski *Never Let a Serious Crisis Go to Waste* 275–286; Verso, 2013). Three decades of social-sciences research on science and politics make it clear that cost–benefit models cannot tame policy-relevant uncertainties or promote political agreement (see, for example, D. Collingridge and C. Reeve *Science Speaks to Power* 3–4, 59–60; Pinter, 1986).

Models that predict higher costs of climate change might make political intervention more palatable. But prescribing models that generate more precisely quantified estimates of a desired output is a political programme, not a scientific one. Responsible research requires responsible quantification and responsible acknowledgement of uncertainty.

Andrea Saltelli* *Autonomous University of Barcelona, Cerdanyola del Vallès, Spain.*
andrea.saltelli@uib.no
**On behalf of 6 correspondents (see go.nature.com/wamqwt for full list).*

Europe must block hornet invasion

Another notable omission from the European Union's list of invasive alien species that are targeted for action is the Asian yellow-legged hornet, *Vespa velutina nigrithorax* (see J. Pergl *et al. Nature* **531,** 173; 2016). Since its arrival in Europe more than a decade ago, this voracious honeybee predator has also caused human deaths from its sting (see K. Monceau *et al. J. Pest. Sci.* **87,** 1–16; 2014).

The hornet's impact is severe in Mediterranean countries, where beekeeping is a crucial source of income. Local beekeepers have their own makeshift eradication methods (such as traps of vinegar with glue), but these also kill important insect pollinators.

The species needs to be officially classified as an invader in all European countries, so that funds can be applied to its study and control. Public campaigns are essential to increase people's awareness and understanding of this threat — for example, regarding the differences between wasp species, many of which are vital for ecosystem functions and services.

We urgently need a coordinated EU plan to control this hornet invasion and to mitigate its potentially serious economic and ecological impacts.

Frederico Santarém* *Research Centre in Biodiversity and Genetic Resources (CIBIO/InBIO), Porto, Portugal.*
fredericosantarem@gmail.com
**On behalf of 5 correspondents (see go.nature.com/uao4qe for full list).*

Software for study design falls short

Online software that can improve the design of animal studies is welcome, but it should not replace specialist advice (see *Nature* **531,** 128; 2016).

Animals are complex biological systems. Their organs and tissues have variable and dynamic functions and morphology in pathophysiological conditions. This complexity calls for a holistic perspective from researchers, who can anticipate and tackle different experimental issues through all phases of a study while aiming for reduction, refinement and replacement in animal use (see www.nc3rs.org.uk/the-3rs).

Such multidisciplinary input also helps to overcome limitations in researchers' scientific scope, experience and skill sets and to improve the quality and interpretation of the results (H. A. Adissu *et al. Dis. Model. Mech.* **7,** 515–524; 2014).

David K. Meyerholz *University of Iowa, Iowa City, USA.*
Alessandra Piersigilli *University of Bern, Bern, Switzerland.*
david-meyerholz@uiowa.edu

Silver lining to irreproducibility

There is room for improvement in how science is done and reported, but something can often be learned from irreproducible experiments. The situation may not be as dire as some headlines imply.

It is crucial to include caveats when citing analyses of reproducibility. For example, an often-quoted 2015 survey of factors that could improve the reproducibility of scientific results (see go.nature.com/yxwgmb) noted that there was a low response rate to the questionnaire, a qualifier that is not always mentioned.

It is important to recognize that researchers cannot control for an unknown variable. Take a web tool for identifying unwanted 'passenger mutations' that could confound analyses of transgenic mice (T. Vanden Berghe *et al. Immunity* **43,** 200–209; 2015). This tool arose from reports of mouse phenotypes that, unbeknown to researchers, depended on unintended mutations. This is an example of a useful resource that enhances our understanding of underlying biological phenomena and results from experiments that might otherwise be branded as irreproducible.

It is in this context that scientific societies are pushing to increase experimental rigour and reporting transparency. For instance, guidelines from the Federation of American Societies for Experimental Biology (go.nature.com/zdf89b) aim to help scientists to meet the reproducibility requirements of research funded by the US National Institutes of Health.

Alyssa Ward *Johns Hopkins University School of Medicine, Baltimore, Maryland, USA.*
Thomas O. Baldwin *University of California, Riverside, USA.*
Parker B. Antin *University of Arizona, Tucson, USA.*
award30@jhmi.edu

Social cooperation among agnostics

Benjamin Grant Purzycki and colleagues suggest that religion helps to explain cooperation in large societies (*Nature* **530,** 327–330; 2016). In my view, knowledge of others' reputations forms a more stable basis for cooperation.

A network with redundant connections transmits these reputations (J. Bruggeman *Social Networks*; Routledge, 2008). It also avoids the strategic manipulation of information by religious entrepreneurs. Once such a cohesive network is established, religious solidarity can enhance cooperation, as can a shared enemy (J.-K. Choi and S. Bowles *Science* **318,** 636–640; 2007) — but it is not essential.

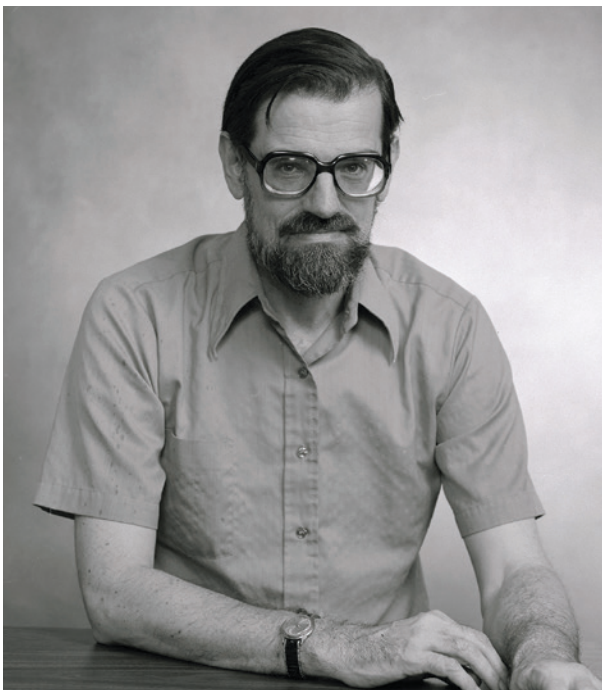
Take the revolt against Communist regimes in 1989. These were overthrown by large-scale collective action, even though religion was negligible or subservient in those countries. Protesters united, despite each knowing only a few others (the regimes suppressed their critics). Religion is one road towards cooperation between strangers, as the experiments show, but not the only one.

Jeroen Bruggeman *University of Amsterdam, the Netherlands.*
j.p.bruggeman@uva.nl

Lloyd Shapley

(1923–2016)

A founding father of game theory.



RAND CORPORATION

Lloyd Stowell Shapley made profound contributions to almost every area of game theory — a field of mathematics that tries to understand how people's choices influence others'. His findings have been applied to all sorts of settings, from politics to hospitals. His whimsically titled 1962 paper — 'On College Admissions and the Stability of Marriage'— published in *The American Mathematical Monthly* and co-authored with the late mathematician David Gale, won Shapley the 2012 Nobel prize in economics, which I shared with him.

Shapley, who died on 12 March, was born in 1923 in Cambridge, Massachusetts, to the astronomer Harlow Shapley and his wife Martha. He pursed a mathematics degree at Harvard University in Cambridge. In 1943, during his third year and at the height of the Second World War, he was drafted into the US Army. In his years of service, he worked at an air base in China and won the Bronze Star, a US military decoration, for breaking a code for Soviet weather reports.

After the war, Shapley graduated from Harvard and worked for two years at the RAND Corporation, which at the time provided research and analysis to the US military. There, he began his work on game theory and came to the attention of the field's founder, John von Neumann.

In 1949, Shapley entered the PhD programme in mathematics at Princeton University in New Jersey — then a hotbed for game theory. There, he overlapped with the mathematician John Nash and the economist Martin Shubik (who would become his long-term collaborator), among many others. Shapley rejoined RAND in 1954, and stayed with the organization for 27 years. In 1981, he moved to the University of California, Los Angeles, where he retired in 2001.

Game theory describes any situation in which the pay-offs that participants receive from their actions are at least partly determined by the actions of other people. Shapley was one of the first to formulate and study the 'core of the game' — the set of outcomes (consequences for everyone in the group) with the property that no coalition of players can do better for themselves by coordinating to produce a different outcome.

Not every game has a core outcome. But for those that do, it often indicates how competition will play out. Shapley's paper with Gale explored this concept in the context of two-sided 'matching games' (D. Gale and L. S. Shapley *Am. Math. Mon.* **69,** 6–15; 1962). In these situations, two sets of players (in the paper, boys and girls seeking marriages, and colleges seeking students and students seeking colleges) have preferences about whom they would like to match with.

In a simple model of one-to-one matching — as applies when each player seeks one spouse, for instance — the core outcomes are those that are stable. After everyone in the game has chosen, there are no pairs (of girls and boys, in the 1960s partnering example) who are not matched to each other but would both prefer to be.

In a note written to Shapley in 1960, Gale asked, "For any pattern of preferences, is it possible to find a stable set of marriages?" Shapley gave his answer in a letter to him: "Let each boy propose to his best girl. Let each girl with several proposals reject all but her favorite, but defer acceptance until she is sure no one better will come her way. The rejected boys then propose to their next-best choices, and so on, until there are no girls with more than one suitor. Marry. The result is stable, since the extramarital liaisons that were previously rejected will be disliked by the girl partners, while all others will be disliked by the boy partners."

Thus was born the 'deferred acceptance algorithm', variants of which are used today to assign US medical graduates to their first jobs, and children to state schools in a growing number of US cities.

Other work from the 1970s by Shapley and the late economist Herbert Scarf on the money-free exchange of indivisible goods ('barter exchange') later helped to organize kidney transplants when donors cannot directly donate to the patient of their choice because of incompatibilities. And starting in the 1990s, Shapley's ideas about two-sided matching and extended barter exchange led to a branch of economic engineering called market design, which seeks to find practical ways to fix broken markets.

In the early 1960s, Shapley and John Milnor (an undergraduate at Princeton when Shapley was a graduate student) initiated the study of 'oceanic games'. In these, there is an 'ocean' of many small players each alone having insignificant influence, so only the actions of people en masse can affect the overall outcome. He later explored these with Robert Aumann, another Nobel economics laureate, in their volume *Values of Non-Atomic Games* (RAND Corporation, 1968).

Although Lloyd and I shared the Nobel prize, we never worked together. But his work was fundamental to my own — for instance, on the practical design of marketplaces. He was a forbidding presence at meetings; I suspect shyness was to blame for his apparent fierceness.

There is a crater on the Moon named Shapley, in honour of Lloyd's astronomer dad. In game theory, Lloyd will likewise be remembered for the mammoth impact he had on the field. ■

Alvin E. Roth *is professor of economics at Stanford University, California, USA. He shared the 2012 Nobel Memorial Prize in Economic Sciences with Lloyd Shapley.*
e-mail: alroth@stanford.edu

ANIMAL BEHAVIOUR

Some begging is actually bragging

A meta-analysis of 143 bird species finds huge variation in parental responses to chicks' begging signals, and shows that parental strategies depend on environmental factors, such as the predictability and quality of food supplies.

DOUGLAS W. MOCK

Within just two weeks of its eggs hatching, each songbird parent delivers around 2,000 prey animals to the nest. The food transforms the cold-blooded, naked and tiny hatchlings into warm-blooded, fully feathered flying machines 20 times their original size. Each arrival of a parent at the nest is met with 'begging' signals. But are these signals actually begging, and if not, what do they indicate?

For a quarter of a century, the predominant answer has been that each chick expresses an honest signal of its own need, and parents, thus informed, help weaklings to feed first. However, writing in *Nature Communications*, Caro *et al.*[1] reveal that this may be an oversimplification. They find that, in many species, parents actively disfavour their weakest offspring so that the family food budget covers others adequately. Thus, although the hatchlings of some species may beg, others brag. Offspring signals to parents are regarded as a test paradigm for ideas about animal communication in general, so Caro and colleagues' results may reset the picture.

The signal-of-need model[2] was proposed in 1991 as a provocative and mathematically elegant variant of an earlier argument devised to explain some species' flamboyant sexual advertisements. According to that 'handicap principle', male self-aggrandizement during courtship is curtailed by the costs of the signal (for example, predation risk limits the size of a peacock's fan), such that only the best males can afford to display such finery.

But when applied to begging, this core logic has to be flipped, such that the costly signals are performed mainly by those least able to afford them. This theory assumes that stronger nestmates desist from displaying the signals. Conceding the spotlight in this manner can be explained by kin selection (in which individuals perform behaviours that are costly to themselves but beneficial to relatives), but only if the indicated returns — in this case, extra nieces and nephews — are forthcoming. This assumption of indirect compensation was overlooked as field and laboratory empiricists flocked to study begging. The literature exploded: before the signal-of-need hypothesis was published, a Web of Science search for 'offspring' and 'begging' shows only a handful of papers; from 1992 on, several hundred have been published, most of which support the hypothesis.

When phenomena are studied with only one hypothesis in mind, standards can slide. Offspring 'need' was originally defined in formal evolutionary terms (an individual's prospects

STEVE SHINN

Figure 1 | Feeding strategies. American coot (*Fulica americana*) parents switch their strategy for feeding their highly mobile and exceptionally colourful chicks midway through the cycle. Initially, parents give food to the nearest chick, which automatically confers a competitive advantage on the larger (earlier-hatched) siblings. But after starvation claims one or two runts, the parents reverse that favouritism — preferentially feeding the weakest survivors and sometimes even attacking the strongest ones[7].

for future reproduction), but such factors are impractical for accurate measurement. Instead, much of the support for signal-of-need came from substituting hunger for need. The problem here is that an offspring can be crammed with food even as it dies of malnutrition, and a robust nestling can be made hungry through brief deprivation[3]. It has been shown for dozens of species[4] that depriving youngsters of food induces escalated begging, but that may not reveal future reproductive potential. Desire is not a synonym for need.

Ironically, the opposite view — 'signal-of-quality', wherein parents generally favour stronger offspring over weaklings — had been proposed a year earlier[5], albeit buried in a long paper. Echoing the advertisement roots of sexual signals, that hypothesis requires no inversion of message and no voluntary abstention. Instead, it proposes that strong offspring are essentially bragging. The signal-of-quality concept also aligns with classic life-history theory[6], in which parents engineer offspring disparities that often facilitate brood reduction, for example by hatching some eggs 1–2 days later than the others. If food availability is unpredictable, competitive mismatches expedite the deferred correction of family size.

Caro *et al.* show that both types of offspring-signalling system may exist in nature, because ecological realities constrain what parents can hope to accomplish. In their meta-analysis, the authors assessed key environmental features and the quality and predictability of food supply for each of 143 species. Variation in environmental quality was scored on the basis of high versus low offspring survival and/or experimental manipulations (additions or subtractions of brood or food), and food predictability was inferred from parental strategies (mainly, whether broods hatched synchronously).

The researchers found that these ecological factors were strongly associated with offspring signalling and within-brood patterns of feeding bias that support two very different parental strategies. If food is relatively predictable, natural selection will favour parents that match family size to the indicated family food budget (creating fewer eggs when food is scarce). In this scenario, survival of the whole brood is the best outcome for everyone, such that a lagging chick should beg more and be fed preferentially, without sibling interference. Conversely, in volatile conditions, parents probably do best by overproducing initially and then pruning later, if necessary, on the basis of offspring size or other physical markers (which devalue the role of behavioural signals). Some species, such as American coots (*Fulica americana*; Fig. 1), actually switch their game mid-cycle, initially letting larger young enjoy their parentally conferred size advantage until brood reduction occurs, and then actively catering to the smallest that remain[7].

By validating pluralism in the explanation for offspring signals, Caro *et al.* encourage further expansion of hypotheses. One to consider is simpler than either signal-of-quality or signal-of-need because it does not require the nestling to possess any 'insider information' about its own long-term prospects, either high (indicating quality) or low (indicating need). Instead, a system could work on the basis of the only 'cryptic' information already known to exist — hunger pangs. In tandem with 'public-domain' cues such as body size, offspring signals might simply answer the mundane but useful question, "Who's ready for another worm?", and thus help parents to make fast allocation decisions. Parents already have knowledge of current food conditions, and for the darker question about who is most expendable, they could rely on visible cues such as size and vigour. This signal-of-hunger hypothesis has strong empirical support[4], and may prove a fine example of Occam's razor — the philosophy that the hypothesis that requires the fewest assumptions is often the most plausible. ■

Douglas W. Mock *is in the Department of Biology, University of Oklahoma, Norman, Oklahoma 73019, USA.*
e-mail: dmock@ou.edu

1. Caro, S. M., Griffin, A. A., Hinde, C. A. & West, S. A. *Nature Commun.* http://dx.doi.org/10.1038/ncomms10985 (2016).
2. Godfray, H. C. J. *Nature* **352,** 328–330 (1991).
3. Price, K., Harvey, H. & Ydenberg, R. *Anim. Behav.* **51,** 421–435 (1996).
4. Mock, D. W., Dugas, M. B. & Strickler, S. A. *Behav. Ecol.* **22,** 909–917 (2011).
5. Grafen, A. *J. Theor. Biol.* **144,** 517–546 (1990).
6. Lack, D. *Ibis* **89,** 302–352 (1947).
7. Shizuka, D. & Lyon, B. E. *Ecol. Lett.* **16,** 315–322 (2013).

This article was published online on 30 March 2016.

CANCER GENOMICS

Hard-to-reach repairs

Two studies find that the molecular machinery that initiates gene transcription prevents repair proteins from accessing DNA, resulting in increased mutation rates at sites of transcription-factor binding. SEE LETTERS P.259 & P.264

EKTA KHURANA

The genetic mutations that lead to cancer are caused by diverse, often poorly understood processes, some of which involve exposure to external agents. Excessive ultraviolet light is linked to melanoma, for example, and tobacco smoke to lung cancer. A molecular mechanism called nucleotide excision repair deals with UV- and smoke-induced genetic damage by removing damaged pieces of DNA, preventing mutations from arising. However, this process is complicated by the fact that repair occurs alongside other crucial genetic activities, such as DNA transcription. Two papers[1,2] in this issue of *Nature* demonstrate how interplay between the DNA-repair and transcription-initiation machinery leads to an increased mutation rate in regulatory regions of the genome.

Although most cancer studies have focused on mutations in protein-coding DNA, there is a growing understanding of the importance of the non-coding DNA regions that regulate gene expression[3–6] — promoter sequences, which are located close to genes, and distant elements called enhancers. Binding of these regions by transcription factors modulates the expression levels of associated genes. On page 264, Sabarinathan *et al.*[1] describe the use of whole-genome sequences from human melanoma samples to analyse mutations in regulatory regions. They found that the cores of the regulatory regions, where transcription factors are predicted to bind, have a mutation rate five times higher than the flanking sequences.

Because of the major role of nucleotide excision repair (NER) in fixing UV-induced DNA damage, Sabarinathan and colleagues next analysed the locations of NER activity[7]. This revealed that the increased mutation rates at transcription-factor binding sites were caused by reduced levels of NER. The authors reasoned that mutations in other cancers that rely on NER should also exhibit this pattern. And indeed, they found increased mutation rates at transcription-factor binding sites in lung-cancer samples, particularly for mutations linked to smoking.

On page 259, Perera *et al.*[2] report the analysis of mutations in regulatory elements in multiple cancer types. They found increased mutation density in the centres of active promoters associated with reduced levels of NER. Moreover, the authors' data suggest that mutation density in regulatory regions is linked not only to transcription-factor binding, but also to the level of transcription initiation.

Thus, two independent studies show that NER at regulatory DNA regions is inhibited by the bound transcription-initiation machinery. This discovery is especially interesting in light of a previous study[8] that showed that

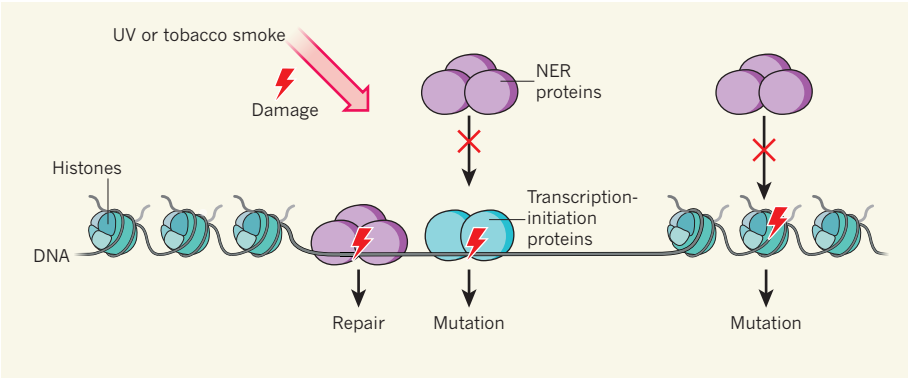


Figure 1 | Easy access prevents mutations. Most DNA is wrapped around histone proteins. By contrast, active regulatory regions are histone-free, to enable binding by transcription-initiation proteins. Exposure to ultraviolet radiation (UV) or tobacco smoke damages DNA, but this damage can be fixed by a process called nucleotide excision repair (NER), which requires DNA binding by NER proteins. Two studies[1,2] now show that NER is disrupted when NER proteins cannot bind DNA because of histones or because of bound transcription-initiation proteins. Mutations accumulate in the inaccessible sites.

mutation density is decreased over active regulatory regions as a whole, relative to their flanking sequences. The authors of that paper proposed that this decrease occurred because active regulatory regions are more accessible than most DNA regions to repair proteins — DNA is typically packaged around proteins called histones, but regulatory regions are unwound for binding by the transcription-initiation machinery. This apparent discrepancy with the current studies reflects the fact that, although regulatory regions as a whole are accessible for NER, the repair machinery is unable to access the core sites within those regions at which transcription factors bind (Fig. 1).

Certain mutations are considered to be drivers of cancer, because they provide a growth advantage to tumour cells. Such mutations are generally identified by the high frequency at which they occur across patients. However, the current studies highlight that protein binding can also lead to high mutation frequency — and so can other factors, such as late replication of a region during cell division[9]. Understanding how these features co-vary with mutation rate is vital for designing accurate computer algorithms to identify driver mutations[10].

It is notable that the variables affecting mutation rate differ for cancer types and subtypes. For instance, unlike in skin and lung cancer, NER does not have a major role in colon cancer. Accordingly, the current studies found no increase in mutation density at the centres of active promoters in colon-cancer samples.

Errors introduced by DNA replication in colon cells are normally resolved by a process called mismatch repair, which is most effective in genomic regions that replicate early during cell division. Thus, mutation rates in colon-cancer cells are generally lower in early-replicating than in late-replicating regions[11]. Mismatch-repair proteins are, however, inactivated in some colon tumours, resulting in the loss of strong correlation between mutation density and replication timing. In fact, the regional 'landscape' of mutation rates can be used to infer the time of mismatch-repair inactivation in the history of a colon tumour.

In the past few years, the complex interplay between DNA-repair mechanisms and genomic properties not originally associated with repair (such as replication timing and DNA accessibility) has become evident, largely thanks to the increasing availability of whole-genome sequences from tumour samples. The need for such sequences from cancer cells has been debated, because they are costly and have limited immediate clinical value[3]. But the current studies demonstrate the immense potential of whole-genome sequences as a lens through which to examine the cellular processes that shape the cancer genome. Genomic studies such as these lay the groundwork for future diagnostic tools and treatments tailored to individuals.

It remains unclear how many more genomic features that correlate with mutation rate are yet to be found. All mutations are ultimately the result of faulty DNA repair — do we need to know all the details of the many ways in which repair can break down to harness the full power of genomics for cancer care? The increasing number of tumour-genome sequences, coupled with our ever-improving knowledge of the machinery involved in genome function, will hold the answer to this question. ■

Ekta Khurana *is at the Meyer Cancer Center and Department of Physiology and Biophysics, Weill Cornell Medical College, New York, New York 10065, USA.*
e-mail: ekk2003@med.cornell.edu

1. Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. *Nature* **532,** 264–267 (2016).
2. Perera, D. *et al. Nature* **532,** 259–263 (2016).
3. Khurana, E. *et al. Nature Rev. Genet.* **17,** 93–108, (2016).
4. Horn, S. *et al. Science* **339,** 959–961 (2013).
5. Huang, F. W. *et al. Science* **339,** 957–959, (2013).
6. Khurana, E. *et al. Science* **342,** 1235587 (2013).
7. Hu, J., Adar, S., Selby, C. P., Lieb, J. D. & Sancar, A. *Genes Dev.* **29,** 948–960 (2015).
8. Polak, P. *et al. Nature Biotechnol.* **32,** 71–75 (2014).
9. Schuster-Böckler, B. & Lehner, B. *Nature* **488,** 504–507 (2012).
10. Lawrence, M. S. *et al. Nature* **499,** 214–218, (2013).
11. Supek, F. & Lehner, B. *Nature* **521**, 81–84 (2015).

REGENERATION

Not everything is scary about a glial scar

After spinal-cord injury, cells called astrocytes form a scar that is thought to block neuronal regeneration. The finding that the scar promotes regrowth of long nerve projections called axons challenges this long-held dogma. SEE ARTICLE P.195

SHANE A. LIDDELOW & BEN A. BARRES

It has long been a mystery why neurons in the peripheral nervous system can regenerate long projections called axons following injury, whereas neurons in the central nervous system (CNS) cannot[1]. One difference is that injured CNS axons lose their intrinsic ability to regrow, but studies have also implicated differences in non-neuronal cells called glia[1,2], which surround neurons to support them and provide insulation. Damaged glia in the CNS release inhibitors of axon regeneration[1], and reactive CNS astrocytes — a type of activated glial cell found at the damaged site — also seem to be powerfully inhibitory[3]. Research[1–3] into spinal-cord injury has centred mostly on the consequences of removing or inhibiting development of the reactive-astrocyte scar. It thus comes as a surprise that Anderson *et al.*[4], on page 195 of this issue, find that this scar in fact strongly supports axon regeneration after spinal-cord injury.

Several studies[5–8] in which reactive

astrocytes were ablated following stroke and spinal-cord injury have shown that the glial scar has a beneficial role in reducing inflammation and secondary tissue damage, and in promoting the recovery of neuronal activity. These findings seem at odds with a previous study indicating that reactive astrocytes inhibit axon regeneration[3] (Fig. 1a). However, this latter study was largely correlational, because it did not directly manipulate reactive astrocytes.

Anderson *et al.* set out to directly test the role of reactive astrocytes in axon regeneration following experimental spinal-cord injury, in which axons die and retract. In their first set of experiments, the authors prevented the formation of reactive astrocytes using two genetically modified mouse models to ablate or attenuate the glial scar — in the first model, proliferating scar-forming astrocytes were selectively killed[5], and in the second, the transcription factor STAT3 (which is required for formation of scar-forming reactive astrocytes) was deleted in astrocytes. In both cases, the researchers found that spontaneous regrowth of damaged axons through the scar did not occur. In addition, animals lacking the scar showed greater dieback of axons from the injury site than was seen in injured wild-type animals, suggesting that reactive astrocytes actually support injured neurons (Fig. 1b).

In a second set of experiments, Anderson *et al.* genetically altered astrocytes to express a receptor for diphtheria toxin, before injuring the spinal cord and allowing a scar to form for five weeks. The researchers then ablated the scar by killing astrocytes with ultralow doses of diphtheria toxin. Again, axons failed to regrow through regions depleted of reactive astrocytes. Moreover, the authors observed pronounced tissue degeneration and found that a larger area contained no axons in the astrocyte-scar-free injuries than under control injury conditions. These data demonstrate that, rather than being detrimental to neuronal health and regenerative capacity, the chronic astrocyte scar is crucial for sustained tissue integrity — upending the long-held dogma.

If reactive astrocytes are inhibitory as previously reported[3], might they exert this effect by altering the extracellular environment to inhibit axon regrowth, for example by upregulating molecules called chondroitin sulfate proteoglycans (CSPGs) that inhibit growth? Anderson *et al.* show that levels of such CSPGs are indeed significantly higher in the injured spinal cord than in the same region of control, uninjured animals. But, surprisingly, when the scar was removed, CSPG levels did not fall, and growth-promoting CSPGs were also upregulated in scar-forming astrocytes. Moreover, aggrecan — a classic growth-inhibitory CSPG used in cell-culture experiments as a measure of reactive-astrocyte inhibition of axon regrowth — was not detected in scar-forming astrocytes. Thus, the glial scar is not the primary source of inhibitory CSPGs, quashing another theory.

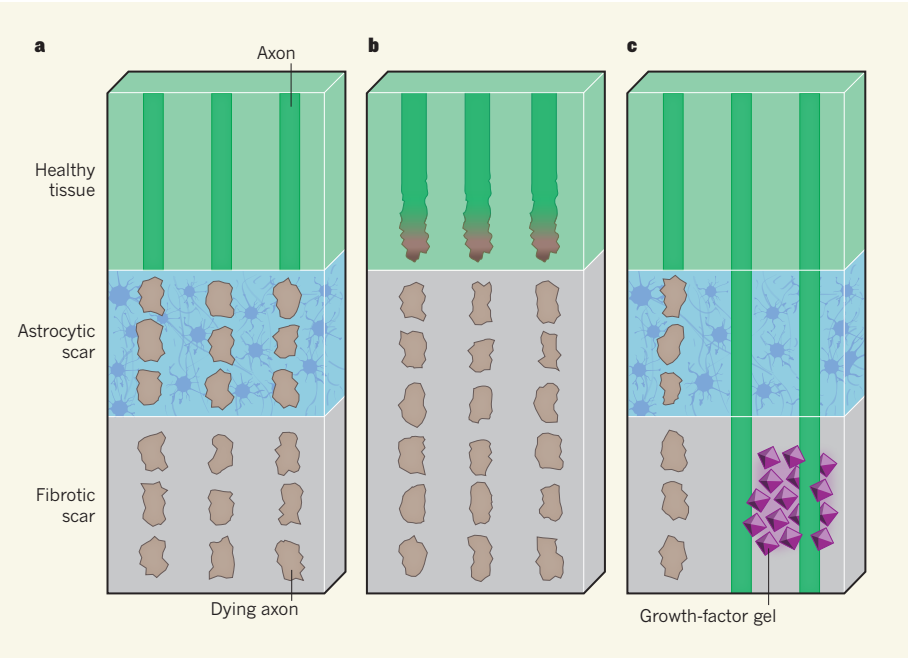


Figure 1 | A model of neuronal regrowth. **a**, Spinal-cord injury damages neurons, which causes long neuronal projections called axons to die back. Injury also activates non-neuronal cells called astrocytes that, along with fibroblast cells, form a scar. It has long been thought that the astrocytic scar inhibits axon regeneration following injury, thus preventing recovery. **b**, However, Anderson *et al.*[4] show that removal of the astrocytic scar does not induce axon regrowth, but instead promotes dieback. **c**, The presence of the scar, when combined with injection of a gel containing growth-promoting factors into the fibrotic scar, actually promotes regrowth in mice.

Axons do not grow by default, but rely on external stimulatory growth cues. Anderson *et al.* next injected a gel containing growth-promoting factors (the neurotrophins NT3 and BDNF) into the injury site to activate neuronal growth programs. This stimulated axons to robustly regrow directly through the dense astrocytic scar tissue (Fig. 1c). In fact, the authors found axons growing along reactive astrocytes. When the scar was removed, neurotrophins alone did not foster axon regrowth. Taken together, these data provide powerful evidence that astrocyte-scar formation aids, rather than inhibits, axon regeneration after injury.

How can these findings be squared with previous studies suggesting that the glial scar is strongly inhibitory? For one thing, other inhibitory cell types, such as fibroblasts and pericytes[9], also contribute to the glial scar. In addition, one study[10] has identified different types of reactive astrocyte. It is therefore possible that, in previous studies, different types of injury produced different types of reactive astrocyte, with some types being inhibitory and others not.

Going forward, it will be important to define the signalling mechanisms that induce activation of the different types of reactive astrocyte. Studies of each cell type should then define their functions, whether they can inhibit axon growth and the molecular mechanisms that underlie their roles. This knowledge could enable the selective manipulation of certain astrocytes by specific molecules, which is preferable to deleting an entire cell population that may well both promote and inhibit axon regrowth. In any case, Anderson and colleagues have shown that, in spite of long-held beliefs to the contrary, reactive astrocytes may not be the villains of spinal-cord recovery, but instead might provide new hope for the regeneration of damaged axons. ■

Shane A. Liddelow *and* **Ben A. Barres** *are in the Department of Neurobiology, Stanford University School of Medicine, Stanford, California 94305-5125, USA.*
e-mails: liddelow@stanford.edu; barres@stanford.edu

1. Silver, J., Schwab, M. E. & Popovich, P. G. *Cold Spring Harb. Perspect. Biol.* **7,** a020602 (2014).
2. Brosius Lutz, A. & Barres, B. A. *Dev. Cell* **28,** 7–17 (2014).
3. Davies, S. J. A. *et al. Nature* **390,** 680–683 (1997).
4. Anderson, M. A. *et al. Nature* **532,** 195–200 (2016).
5. Bush, T. G. *et al. Neuron* **23,** 297–308 (1999).
6. Faulkner, J. R. *et al. J. Neurosci.* **24,** 2143–2155 (2004).
7. Okada, S. *et al. Nature Med.* **12,** 829–834 (2006).
8. Herrmann, J. E. *et al. J. Neurosci.* **28,** 7231–7243 (2008).
9. Göritz, C. *et al. Science* **333,** 238–242 (2011).
10. Zamanian, J. L. *et al. J. Neurosci.* **32,** 6391–6410 (2012).

This article was published online on 30 March 2016.

PHYSICS

Quantum problems solved through games

Humans are better than computers at performing certain tasks because of their intuition and superior visual processing. Video games are now being used to channel these abilities to solve problems in quantum physics. SEE LETTER P.210

SABRINA MANISCALCO

In March, the mobile-game developer Supercell released a video thanking the 100 million people around the world who play its games daily[1]. According to estimates, by the age of 21 the average US citizen has spent more than 10,000 hours playing video games[2] — the equivalent of working in a full-time job of 40 hours per week for five years. So is it possible to channel the enormous amount of human brain-power used in this way by designing games that have a purpose? More specifically, can video games be developed in which people solve computationally intractable research problems as a side effect of playing? Writing on page 210 of this issue, Sørensen *et al.*[3] answer this question with a resounding 'yes'. They have developed video games that help to solve a problem relating to the realization of a new, efficient and scalable architecture for a quantum computer.

Despite their capacity to handle vast amounts of data, there are many problems in science that computers cannot yet solve. It is therefore highly desirable to harness innate human abilities to perform tasks that are beyond the grasp of current machines. The challenge is how to turn a research problem into a game, a process known as gamification.

Such a game should not only have a structure that embeds the specific research problem — with rules and winning conditions that encourage exploration of the 'landscape' of solutions — but also be fun to play. Successful examples include citizen-science games such as *Foldit* (ref. 4), *EteRNA* (ref. 5) and *EyeWire* (ref. 6), which are used to study protein folding, RNA folding and neuron mapping, respectively. Their success stems from humans' intuition and superior understanding of the 'real' world.

In this context, Sørensen and colleagues' work is an amazing feat, because quantum mechanics is the most counterintuitive and bizarre of all physical theories. Quantum particles behave in several unusual ways. They can be in several locations at the same time, can tunnel through potential-energy barriers, and can correlate with each other in such a way that their individual identity is lost. Niels Bohr's quote, "Anyone who is not shocked by quantum theory has not understood it", is as relevant now as it was when quantum physics was first developed some 100 years ago.

The authors developed a video game called *Quantum Moves* (ref. 7) in an effort to solve one of their own practical problems. They are working on a prototype of a quantum computer based on atoms trapped in an artificial crystal of light[8] — periodic potential-energy wells generated by lasers. Quantum logic-gate operations can be performed in this system by moving atoms using a highly focused laser beam (an optical tweezer)[9].

But quantum-computing operations must be executed fast — faster than the time taken for the quantum state of a system to lose its quantum properties[10]. This is where the real difficulty arises: how can the optimal quantum move be found in the shortest possible time? Moreover, for perfect fidelity of gate operations, there is a fundamental limit to the minimum duration of each operation, known as the quantum speed limit. Numerical approaches have failed to accurately predict the quantum speed limit for the class of time-dependent problems tackled by *Quantum Moves*.

One of the challenges of *Quantum Moves* is called BringHomeWater. In this challenge, players have to move an optical tweezer to a region where an atom is trapped in a simplified version of the artificial light crystal. They then have to collect the atom and move it to a target area as fast as they can — simulating one of the basic steps needed to operate a logic gate or perform a quantum simulation in the authors' prototype quantum computer. The atom is visualized as a quantum-mechanical wavefunction, which looks like water in a glass (Fig. 1). The faster the atom is moved, the easier it is to 'spill' the water. The players thus have to find the fastest way to bring home the atom without losing it (spilling the water) along the way, providing information that helps the researchers to optimize atom movements in the quantum computer.

Sørensen *et al.* show that move-optimization schemes that use the players' solutions outperform the best known strategies devised by computers, and provide new lower bounds to the quantum speed limit. The players have thus identified the 'settings' needed for the optimal implementation of simple quantum-logic operations. Not only that, their solutions also helped the researchers to understand the physical mechanisms by which such strategies work in the weird quantum world.

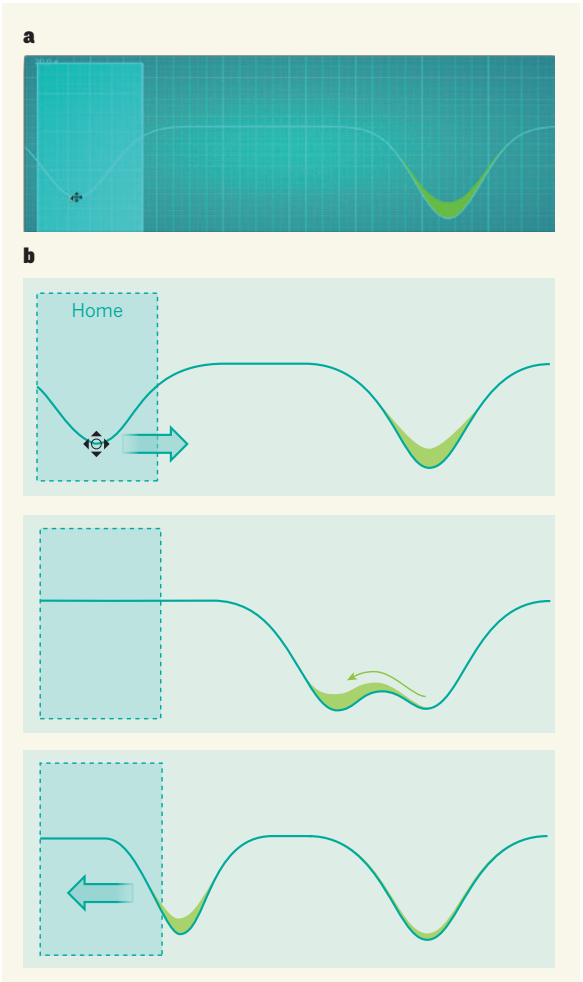


Figure 1 | The online game BringHomeWater. **a**, This snapshot shows a still image of BringHomeWater, part of a game developed by Sørensen and colleagues[3]. **b**, In the game, players move an optical tweezer (represented by the left-hand curve; the black symbol is a cursor used to move the tweezer) from the home region towards an atom (depicted as a light-green liquid) in a potential-energy well (the right-hand curve). The players must find a route that brings the liquid home as quickly as possible, without spilling it. Sørensen *et al.* used the fastest moves found by players to optimize the transfer of atoms within their quantum computer; such optimization is required for quantum-logic operations. Large arrows show tweezer movements.

So how can game players solve difficult research problems in quantum theory when they have no knowledge of either the puzzling phenomena of quantum physics or the sophisticated mathematical formalism used to describe it? One can do things in games that cannot be done in reality, so gamers are used to experimenting with possibilities that go beyond the classical laws of physics. Perhaps this ability to think outside the box allows them to make the creative leap necessary to tackle quantum problems.

Understanding the principles and key conditions for the successful gamification of quantum problems is an interdisciplinary endeavour requiring the interaction and collaboration of quantum physicists, game researchers, neuroscientists and many others. Whether Sørensen and colleagues' method will be applicable to a wide range of problems in quantum physics is currently an open question. But because we are on the verge of a new era of quantum technologies, this approach is definitely worth pursuing, and is a theme of initiatives such as *Quantum Game Jam*[11]. ■

Sabrina Maniscalco *is at the Turku Centre for Quantum Physics, Department of Physics and Astronomy, University of Turku, Turku 20014, Finland.*
e-mail: sabrina.maniscalco@utu.fi

1. https://www.youtube.com/watch?v=0sQAj9p_ppU
2. von Ahn, L. & Dabbish, L. *Commun. ACM* **51,** 58–67 (2008).
3. Sørensen, J. J. W. H. *et al. Nature* **532,** 210–213 (2016).
4. Cooper, S. *et al. Nature* **466,** 756–760 (2010).
5. Lee, J. *et al. Proc. Natl Acad. Sci. USA* **111,** 2122–2127 (2014).
6. Kim, J. S. *et al. Nature* **509,** 331–336 (2014).
7. https://www.scienceathome.org/quantum-moves/game
8. Bloch, I. *Nature Phys.* **1,** 23–30 (2005).
9. Weitenberg, C., Kuhr, S., Mølmar, K. & Sherson, J. *Phys. Rev. A* **84,** 032322 (2011).
10. Zurek, W. H. *Phys. Today* **44,** 36–44 (1991).
11. http://www.quantumgamejam.com

NEUROINFLAMMATION

Surprises from the sanitary engineers

In mammals, microglial cells of the central nervous system are responsible for the normal clearance of dead brain cells. TAM–receptor proteins have now been found to mediate this function. SEE LETTER P.240

RICHARD M. RANSOHOFF

Cell-surface receptors are specialized molecules that respond to precise signals, so that environmental input elicits commensurate responses. On page 240 of this issue, Fourgeaud *et al.*[1] describe how they manipulated the mouse genome to delete receptor proteins of the TAM family from microglia — a type of brain cell distantly related to the resident inflammatory cells found in tissues such as the skin, spleen and liver. The results provide startling insight into the process by which the adult brain generates new neurons, and open up avenues for studying microglia.

TAM receptors (named from the first letters of the member proteins Tyro3, Axl and Mer) are evolutionarily recent, appearing first in the invertebrate sea squirts. Newly emerged gene families often have highly refined roles, and their function is commonly dispensable for embryonic development. People (or genetically engineered mice) with defective TAM genes develop normally but show varied effects later in life[2]. For example, humans or rodents deficient in the *Mer* gene develop a form of retinitis pigmentosa. This degenerative eye disease occurs because rod photoreceptor cells (RPCs) accumulate toxic by-products of chemical reactions through which light is converted to nerve impulses. This waste material is removed through engulfment of the RPCs' outer segment by retinal pigment epithelial cells; without Mer, this process fails and RPCs die.

Another example is that mice lacking all three TAM receptors show male infertility, because vast numbers of superfluous germ cells die and accumulate in the testes, leading to degeneration of the remaining, otherwise-viable germ cells. Mice that lack individual receptors or ligands of the TAM system also show blood-clotting defects, and those deficient in all three receptors develop widespread autoimmune responses that are reminiscent of the human disease systemic lupus erythematosus[3]. Thus, TAM-receptor signalling is used in a wide variety of disparate functions primarily associated with the removal of dying cells and waste material.

Mice that lack all three TAM receptors are normal at birth, indicating that these proteins are not required to eliminate dead cells during embryonic development[2]. But the TAM receptors are involved in a variant form of this elimination mechanism to achieve dynamic tissue remodelling throughout life[2]. This cell-corpse

50 Years Ago

Engineers get the rough end of the stick even in countries where they are more esteemed than in ours … You fire off a rocket and a satellite moves successfully around the world. That is a scientific triumph. On the other hand, if it flops on the launching pad, that is an engineering failure … It is no accident that among our scientists there is still a cheerful and relaxed attitude to qualifications. Cockcroft is a great physicist, but he has never taken a physics course in his life … Crick has revolutionized modern biology; but he has had as much formal instruction in biology as he has in Hebrew … Engineering has … suffered through the rigidity of its training … Where the esteem and the rewards appear to be, there able people will go … Contemporary engineering education does not encourage enough the speculative and rebellious intelligence … It is rare for engineering students to question everything under heaven or earth in the way that good scientific students will … If we get our education right … the place of the engineer in society will become right … To most sane persons, esteem is more important than pay. If we had a choice most of us, I hope, would prefer to be President of the Royal Society than the most successful pop singer in the world.

Lord Snow

From *Nature* 16 April 1966

100 Years Ago

The old Romans and Greeks, as evidenced by the statues, were evidently gentlemen addicted to shaving, but … the means of producing soap in those days must have been limited. The only conclusion that one can arrive at is that they must have shaved without soap.

From *Nature* 13 April 1916

removal, as well as that occurring in the wake of an infection, is carried out in vertebrates by phagocytes ('cells that eat'), and particularly by white blood cells termed macrophages ('big eaters'). Cells that are undergoing apoptotic cell death expose surface markers ('eat me' signals) that alert nearby phagocytes to engulf and dispose of the cell corpse. But, in contrast to receptors that directly bind cells displaying these markers, TAM receptors use adaptor proteins, or ligands, that bind both the marker on the cell to be removed and the TAM receptor on the phagocyte. These adaptors are called growth-arrest-specific protein 6 (Gas6) and protein S.

Fourgeaud *et al.* examined the function of TAM receptors in brain microglia — cells whose origin and competencies are only now becoming clear. Microglia are derived from primitive macrophages that enter the embryo from the yolk sac early in development, become distributed throughout the embryo as resident macrophages, and guide organ development. In the embryonic brain, which is populated by microglia from day 10 of gestation in mice and from gestational week 4.5 in humans[4], exuberant production of redundant cells keeps microglia busy clearing corpses[5]. The brain forms normally without TAM receptors, and Fourgeaud and colleagues wondered what the molecules' roles might be in adult life.

The study initially focused on two brain regions in which neurons are continuously being born and that are therefore designated neurogenic niches[6]. One of these niches produces neurons to replenish olfactory neurons, which support the sense of smell. Neurons from the second niche become integrated into regions associated with memory and learning. As with many generative tissues, the neurogenic niches produce an excess of progenitor cells (which have the potential to develop into neurons), most of which die. Speculation held that microglia cleared these cell corpses, and it seemed plausible that neurogenesis would fail if corpse removal was impaired[7]. Previous research[8] using mice that lacked all three TAM receptors in all tissues suggested an alternative idea: that neurogenesis is suppressed as a result of excessive inflammatory reactions by microglia with deficient TAM signalling. However, the state of the neurogenic niches without microglial TAM receptors remained unresolved.

Fourgeaud and colleagues' investigation of mice that lacked both *Mer* and *Axl* initially confirmed the predicted phagocytic defect: neuron progenitors showing markers of apoptotic death accumulated to a striking degree, whereas these dying cells were not seen in wild-type mice, nor were they seen in regions other than the niches. But the authors' analysis of neurogenesis yielded a shock: new neurons in the olfactory region increased by 70% in the mice lacking *Mer* and *Axl*.

An explanation for this result came from another study[9], which showed that microglia

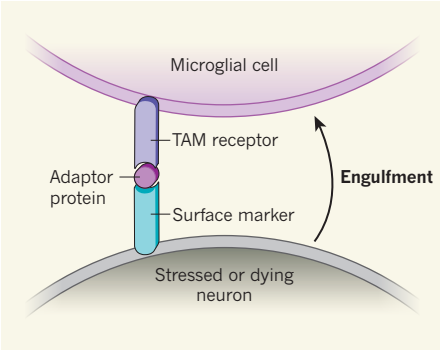


Figure 1 | Neuron clearance. Fourgeaud *et al.*[1] show that TAM-receptor proteins are required for microglial cells in the brain to remove neural cells that exhibit surface markers indicative of stress or apoptotic cell death. TAM-receptor binding, which occurs through an adaptor protein, leads to the engulfment of the stressed or dying cell by the microglial cell.

in regions of the mouse brain damaged by stroke engulfed viable neurons that displayed 'eat me' signals because of stress, not cell death. This study also found that if the uptake of these neurons by phagocytes was blocked, and thus their death by 'phagoptosis' prevented, the severity of stroke was reduced. Fourgeaud and colleagues' interpretation of these combined data is that phagoptosis can be observed in the healthy brain during neurogenesis (Fig. 1). Future work to determine the effects of augmented neurogenesis will be informative.

The *Mer*- and *Axl*-deficient microglia showed other changes as well. Microglia are noted for their delicate, branched processes, which are continually in motion, monitoring synapses (connections between nerve cells)[10]. Fourgeaud *et al.* show that TAM receptors are essential for the processes to be fully motile. Would lack of one or more of these receptors change microglial cells' ability to carry out their manifold tasks in aid of synaptic networks? Further revelations are likely as the TAM-receptor story unfolds and is integrated with findings from other microglial signalling systems. ■

Richard M. Ransohoff *is in the Neuroimmunology group, Biogen, Cambridge, Massachusetts 02142, USA.*
e-mail: richard.ransohoff@biogen.com

1. Fourgeaud, L. *et al. Nature* **532,** 240–244 (2016).
2. Lemke, G. *Cold Spring Harb. Perspect. Biol.* **5,** a009076 (2013).
3. van der Meer, J. H., van der Poll, T. & van 't Veer, C. *Blood* **123,** 2460–2469 (2014).
4. Verney, C., Monier, A., Fallet-Bianco, C. & Gressens, P. *J. Anat.* **217,** 436–448 (2010).
5. Gomez Perdiguero, E., Schulz, C. & Geissmann, F. *Glia* **61,** 112–120 (2013).
6. Aimone, J. B., Deng, W. & Gage, F. H. *Trends Cogn. Sci.* **14,** 325–337 (2010).
7. Sierra, A. *et al. Neural Plastic.* **2014,** 610343 (2014).
8. Ji, R. *et al. J. Immunol.* **191,** 6165–6177 (2013).
9. Brown, G. C. & Neher, J. J. *Trends Biochem. Sci.* **37,** 325–332 (2012).
10. Sierra, A., Tremblay, M.-E. & Wake, H. *Front. Cell. Neurosci.* **8,** 240 (2014).

This article was published online on 6 April 2016.

GEOCHEMISTRY

How rain affects rock and rivers

An analysis of the evolution of river channels on Hawaii's Big Island shows that a key factor is the effect of local rainfall on bedrock strength — rather than its effect on river discharge, as is often assumed. SEE LETTER P.223

ALISON M. ANDERS

On page 223 of this issue, Murphy *et al.*[1] report that rainfall has a marked tendency to weaken rock through chemical weathering, so that increases in rainfall raise local rates of river erosion. This finding is valuable because it helps to provide a basis for quantifying the relationship between chemical weathering, which consumes atmospheric carbon dioxide, and erosion by rivers, which sets the pace of landscape-wide erosion in unglaciated mountain ranges.

The global carbon cycle is influenced by a range of processes that occur over geological timescales of tens of millions of years[2]. For example, large quantities of carbon can be stored underground in coal, oil and limestone as sedimentary basins evolve, and atmospheric CO_2 is absorbed when mountain building exposes silicate minerals to chemical-weathering reactions (Fig. 1). On the other side of the balance sheet, carbon stored in the deep Earth is reintroduced to the atmosphere by volcanism and metamorphism (changes in the mineral content and structure of rocks that occur at moderate pressures and temperatures, excluding melting).

Potential feedbacks in this long-term carbon cycle are embodied in the relationships between the rates of: mountain building; chemical and physical erosion of mountainous

topography; drawdown of atmospheric CO_2; and changes in global climate. Some of these relationships have been debated[3–5] as potential explanations for the co-occurrence over the past 50 million years of profound global cooling, the rapid growth of mountain ranges, increasing dominance of glacial erosion in shaping mountain ranges and (until the Industrial Revolution) declining atmospheric CO_2 levels. An understanding of the impact of climate on erosion rates is central to unravelling these complex relationships. However, developing this understanding has proved difficult[6,7].

To address this problem, Murphy *et al.* measured the variability in strength and chemical composition of basalt (the most common volcanic rock) along river valleys that cross large gradients in mean annual precipitation on the Big Island of Hawaii. The topography of the Big Island before any erosion took place is known, because the island formed from a 'shield' volcano, and shield volcanoes have predictable geometries. The topographic profiles along rivers on the Big Island can therefore be used to derive average rates of river incision into rock over the well constrained age of the basaltic lava flows that formed the island.

The authors compared these river-incision rates with modern mean annual precipitation and rock strength. On the dry side of the island, they found that rock strength decreases as modern mean annual precipitation increases. This pattern reveals the decreasing strength of basalt as it undergoes progressive chemical weathering. But a different relationship holds on the wet side of the island: rock strength is lower than on the dry side and increases with increasing river-incision rate. This is because chemical weathering has generally progressed to a more advanced stage than on the dry side, and the exposure of fresh rock by river incision increases rock strength where incision rates are large.

Murphy *et al.* then developed a predictive model of river incision that included the effect of local precipitation on bedrock strength, and show that it reproduces the observed river-channel topography. The model cannot reproduce the observations when this effect is neglected, even when the spatial variability in precipitation is accounted for in simulations of river discharge (rate of flow). The model suggests that chemical weathering becomes a substantial modulator of physical erosion rates where such erosion rates are modest, but that the impact of chemical weathering on overall erosion is much less where physical erosion is rapid.

Although the complementary nature of chemical weathering and physical erosion has long been recognized conceptually, this work quantifies it in the context of landscape evolution — something that has seldom been done[8]. But Murphy and co-workers' study also highlights a subtle, crucial and frequently overlooked assumption that has hindered our understanding of the connection between climate and erosion: the idea that the effect of spatial variability in precipitation is dominated by the influence of precipitation on river discharge. Instead, the authors show that the main effect of precipitation is its role in driving chemical weathering. Spatial variability in precipitation does strongly influence spatial patterns of resistance to erosion observed in river-channel topography. But the impact of precipitation gradients on river discharge is much less pronounced, because discharge depends only on the average precipitation across the whole drainage area of the river.

The finding that the major influence of spatial variability in precipitation on river erosion is through chemical weathering also suggests other ways in which precipitation might influence local erosion rates. For example, precipitation controls the type of vegetation that can grow, which in turn influences chemical-weathering rates and physical resistance to erosion.

Although the authors' work to quantify the influence of precipitation on chemical weathering and rock hardness is admirable, much uncertainty remains. Precipitation patterns were probably stable over the several hundreds of thousands of years during which the river channels evolved, but we have few constraints on estimates of the absolute amount of precipitation at times other than the past century. Moreover, for most of this time period, the global climate was cooler than it is now, suggesting that modern mean annual precipitation rates in Hawaii are probably higher than the average of such rates over the lifetime of the channels[9]. This implies that the reported sensitivity to precipitation may be an underestimate, and highlights the need for caution in applying the observed relationship more broadly.

Nevertheless, Murphy and colleagues' study is an invaluable contribution to our growing understanding of interactions between global climate, rock weathering, mountain erosion and the long-term carbon cycle. Tropical volcanic islands will probably contribute disproportionately to global chemical-weathering fluxes because of their warm and wet climates and their supply of fresh, easily weathered basalt[10]. To determine the impact of climate on weathering and erosion globally, however, we also need to understand bedrock-weathering processes in glaciated mountains and cold climates, because these processes are more characteristic of those that occurred over large areas of continental crust in Earth's recent geological history. ■

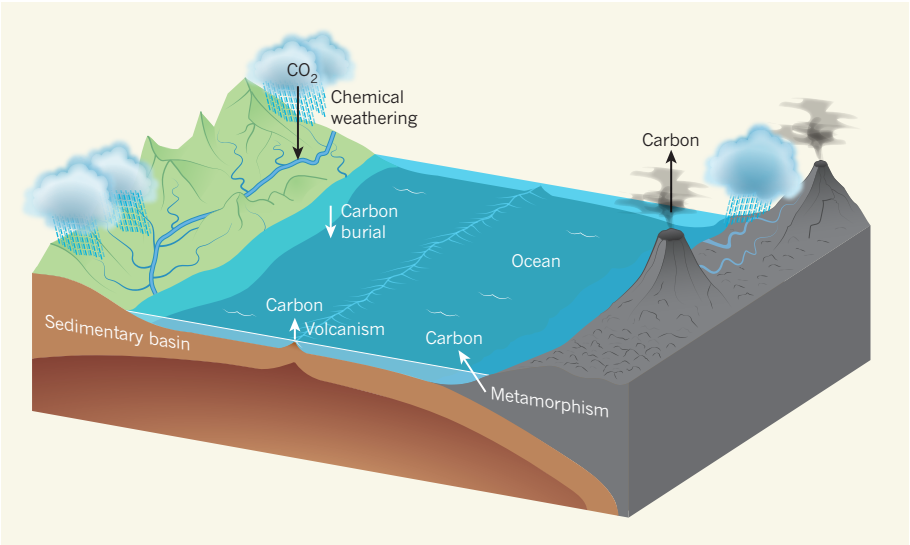


Figure 1 | The geological carbon cycle. On timescales of tens to hundreds of millions of years, carbon cycles between the atmosphere and the deep Earth. Chemical-weathering processes associated with rainfall consume atmospheric carbon dioxide, and carbon also becomes buried in sedimentary basins. Volcanism and metamorphic processes (by which minerals change form without melting) release deep-Earth carbon back into the atmosphere. Murphy *et al.*[1] studied the evolution of river channels on Hawaii's Big Island, and find that local precipitation affects bedrock strength and chemical weathering — improving our understanding of the relationships between the processes that affect the long-term carbon cycle.

Alison M. Anders *is in the Department of Geology, University of Illinois, Champaign, Illinois 61820, USA.*
e-mail: amanders@illinois.edu

1. Murphy, B. P., Johnson, J. P. L., Gasparini, N. M. & Sklar, L. S. *Nature* **532,** 223–227 (2016).
2. Berner, R. A. *Nature* **426,** 323–326 (2003).
3. Molnar, P. & England, P. *Nature* **346,** 29–34 (1990).
4. Whipple, K. X. & Tucker, G. E. *J. Geophys. Res.* **104,** 17661–17674 (1999).
5. Willenbring, J. K. & Jerolmack, D. J. *Terra Nova* **28,** 11–18 (2016).
6. Herman, F. *et al. Nature* **504,** 423–426 (2013).
7. Whipple, K. X. *Science* **346,** 918–919 (2014).
8. Dixon, J. L., Heimsath, A. M. & Amundson, R. *Earth Surf. Process. Landforms* **34,** 1507–1521 (2009).
9. Sheldon, N. D. *J. Geol.* **114,** 367–376 (2006).
10. Gaillardet, J., Dupré, B., Louvat, P. & Allègre, C. J. *Chem. Geol.* **159,** 3–30 (1999).

ARTICLE

doi:10.1038/nature17410

Hourglass fermions

Zhijun Wang[1]*, A. Alexandradinata[1,2]*, R. J. Cava[3] & B. Andrei Bernevig[1]

Spatial symmetries in crystals may be distinguished by whether they preserve the spatial origin. Here we study spatial symmetries that translate the origin by a fraction of the lattice period, and find that these non-symmorphic symmetries protect an exotic surface fermion whose dispersion relation is shaped like an hourglass; surface bands connect one hourglass to the next in an unbreakable zigzag pattern. These 'hourglass' fermions are formed in the large-gap insulators, KHgX (X = As, Sb, Bi), which we propose as the first material class whose band topology relies on non-symmorphic symmetries. Besides the hourglass fermion, another surface of KHgX manifests a three-dimensional generalization of the quantum spin Hall effect, which has previously been observed only in two-dimensional crystals. To describe the bulk topology of non-symmorphic crystals, we propose a non-Abelian generalization of the geometric theory of polarization. Our non-trivial topology originates from an inversion of the rotational quantum numbers, which we propose as a criterion in the search for topological materials.

Spatial symmetries are ubiquitous in crystals. A basic geometric property that distinguishes these symmetries concerns the ways in which they transform the spatial origin: rotations, inversions and reflections preserve the origin, whereas screw rotations and glide reflections unavoidably translate the origin by a fraction of the lattice period[1]. If no origin exists that is simultaneously preserved (modulo lattice translations) by all symmetries in a space group, then this space group is called non-symmorphic. Despite there being more non-symmorphic than symmorphic space groups, a non-symmorphic insulator with non-trivial topology has yet to be found.

We describe a topology in band insulators that arises from fractional translations of the origin, and propose KHgX (X = As, Sb, Bi) as the first material realization of its kind. The topology of KHgX manifests differently on its various surfaces, depending on the spatial symmetries that are preserved on that surface. On the 010 surface, we find that the glide–mirror symmetry protects a surface fermion whose dispersion relation is shaped like an hourglass (see Fig. 1d); doubly-degenerate surface bands connect one hourglass to the next in a zigzag pattern that robustly interpolates across the conduction gap in Fig. 1a. This hourglass fermion sharply contrasts with the Dirac fermions found on the surface of symmorphic topological insulators[2]. The 100 surface of KHgX uniquely realizes a 3D, doubled quantum spin Hall effect (QSHE) with four counter-propagating surface modes distinguished by spin, as illustrated in Fig. 1f. Unlike the well-known 2D QSHE[2–8] (Fig. 1e), the surface states of KHgX are not protected by time-reversal symmetry alone, but are further stabilized by spatial symmetries. To describe the bulk topology of KHgX, we introduce a non-Abelian generalization of polarization that naturally describes glide-symmetric crystals, and is further quantized owing to space–time inversion symmetry. Our work extends the well-known Abelian theory of polarization[9–11], which exhibits quantization due to spatial-inversion symmetry[12].

Crystal structure

The crystal structure of KHgX is illustrated in Fig. 2: Hg and X ions form honeycomb layers with AB stacking along $\boldsymbol{z}$; between each AB bilayer sits a triangular lattice of K ions. The spatial symmetries include: an inversion ($\mathcal{I}$) centred around a K ion, which we choose as our spatial origin; a screw rotation $\overline{C}_{6z}$, which is a sixfold rotation about $\boldsymbol{z}$ followed by a fractional lattice translation ($t(c\boldsymbol{z}/2)$, where c is a lattice constant illustrated in Fig. 2, and the reflections $M_y : (x, y, z) \to (x, -y, z)$, $\overline{M}_z = t(c\boldsymbol{z}/2)M_z$ and $\overline{M}_x = t(c\boldsymbol{z}/2)M_x$. (Here and henceforth, for any transformation g, we denote $\overline{g} = t(c\boldsymbol{z}/2)g$ as a product of g with the fractional translation.) Among the reflections, only $\overline{M}_x$ is a glide reflection, for which the fractional translation is unremovable by a different choice of origin. Altogether, these symmetries generate the non-symmorphic space group $D^4_{6h}(P6_3/mmc)$ (ref. 13).

Each topological feature of KHgX may be attributed to a smaller subset of the group—on surfaces where certain bulk symmetries are lost, their associated topology is not manifest; for example, the 100-surface symmetry is a symmorphic subgroup of D^4_{6h}, leading to a strikingly different band structure to that of the non-symmorphic 010 surface. Our strategy is to deduce the possible topologies of the surface bands purely from representations of the surface symmetry. We then more carefully account for the bulk symmetries and their representations, and introduce a non-Abelian polarization to diagnose non-trivial topology in the bulk wavefunctions.

Surface analysis

Let us first discuss the 010 surface, whose space group ($Pma2$) is generated by glideless $\overline{M}_z$ and glide $\overline{M}_x$. To explain the robust surface bands in Fig. 1, we consider each high-symmetry line in turn.

(1) At any wavevector ($\boldsymbol{k}'$) along $\widetilde{\mathrm{Z}}\widetilde{\mathrm{U}}$, where the z-component of k' is $k_z = \pi/c$, all bands are doubly-degenerate. Indeed, the group[14] of $\boldsymbol{k}'$ includes the anti-unitary element $T\overline{M}_x$ (time reversal with a glide), which results in a Kramers-like degeneracy at each $\boldsymbol{k}'$. This follows from $(T\overline{M}_x)^2 = T^2\overline{M}_x^2 = t(c\boldsymbol{z})$, for which the lattice translation is represented by Bloch waves along $\widetilde{\mathrm{Z}}\widetilde{\mathrm{U}}$ as a phase factor: $t(c\boldsymbol{z}) = \exp(-ik_z/c) = -1$.

(2) Along both glide-invariant lines ($\widetilde{\Gamma}\widetilde{\mathrm{Z}}$ ($k_x = 0$) and $\widetilde{\mathrm{X}}\widetilde{\mathrm{U}}$ ($k_x = \pi/\sqrt{3}\,a$)), bands split into quadruplets that each exhibits an internal partner-switching in the interval $k_z \in [0, \pi/c]$. To explain, $\overline{M}_x^2 = t(c\boldsymbol{z})\overline{E}$, with $\overline{E}$ a 2π-spin rotation, implies two branches for the mirror eigenvalues: $\pm i\exp(-ik_zc/2)$. The role of time-reversal symmetry is to enforce degeneracies between complex-conjugate representations at both Kramers points; that is, the $\overline{M}_x$ eigenvalues are paired as $\{+i, -i\}$ at $k_z = 0$, and either $\{+1, -1\}$ or $\{-1, -1\}$ at $k_z = \pi/c$. These constraints imply two topologically distinct connectivities for the surface bands. In the first (Fig. 3c), surface bands zigzag across the

[1]Department of Physics, Princeton University, Princeton, New Jersey 08544, USA. [2]Department of Physics, Yale University, New Haven, Connecticut 06520, USA. [3]Department of Chemistry, Princeton University, Princeton, New Jersey 08540, USA.
*These authors contributed equally to this work.

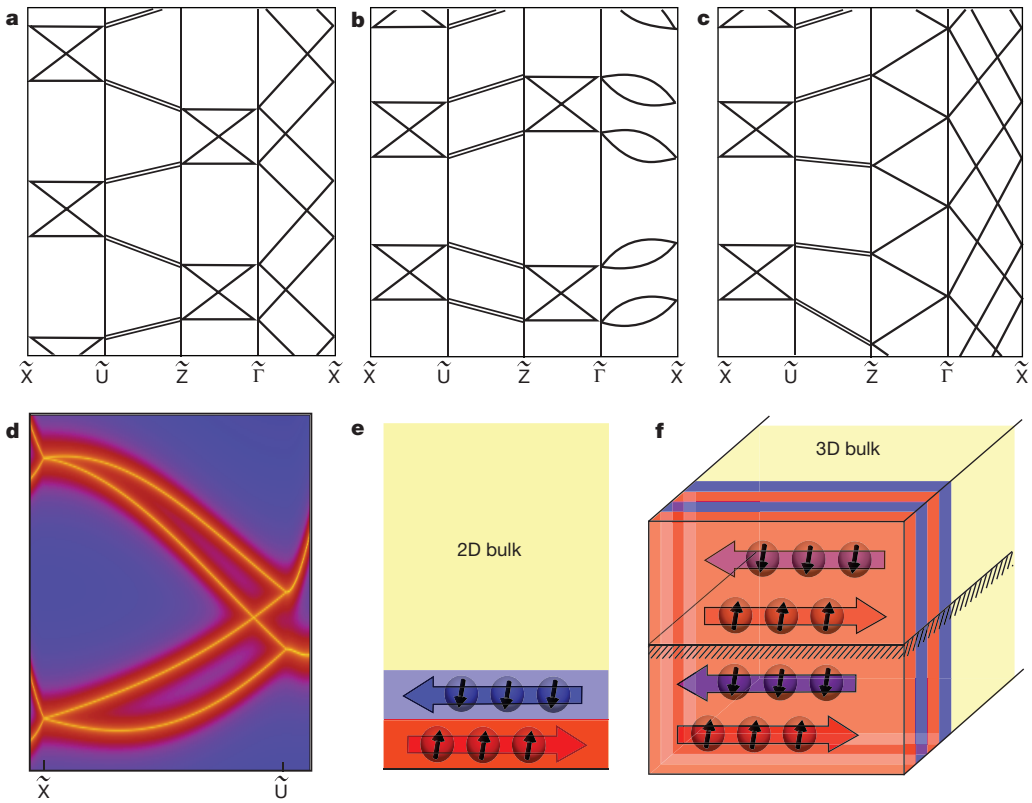


Figure 1 | Hourglass fermions and the 3D quantum spin Hall effect (QSHE). a–c, Examples of the possible topologies of surface bands in a non-symmorphic crystal: our material class (**a**), the trivial topology (**b**) and a non-trivial topology that may be found in other materials (**c**). All crossings along $\widetilde{Z}\widetilde{\Gamma}\widetilde{X}\widetilde{U}$ and degeneracies along $\widetilde{U}\widetilde{Z}$ arise from symmetry. **d**, Orange lines show the dispersion of the hourglass fermion in KHgSb. The vertical axes in **a–d** show energy. **e**, The well-known 2D QSHE, with a pair of spin-split modes, coloured red and blue, counter-propagating on the edge. **f**, Schematic of a 3D, doubled QSHE: two right-going surface modes (red) with spin up and two left-going modes with spin down (blue). These surface modes are protected by a reflection symmetry that is indicated by parallel diagonal lines.

conduction gap and each cusp is a Kramers doublet—this will be elaborated as a glide-symmetric analogue of the 2D QSHE[15]. The second connectivity in Fig. 3d applies to our material class: an internal partner-switching occurs within each quadruplet, resulting in an hourglass-shaped dispersion. The centre of each hourglass is a robust crossing between orthogonal mirror branches; this unremovable crossing is movable along the interval $k_z \in [0, \pi/c]$, as exemplified by KHgSb in Fig. 1d.

Piecing together (1) and (2) along the bent line $\widetilde{X}\widetilde{U}\widetilde{Z}\widetilde{\Gamma}$, we show how a robust interpolation across the energy gap may arise. At $\widetilde{Z}$ and $\widetilde{U}$ there are two ways to connect hourglasses to degenerate doublets: an 'hourglass flow' describes the spectral connection of all hourglasses by zigzag-connecting doublets, as drawn in the $\widetilde{X}\widetilde{U}\widetilde{Z}\widetilde{\Gamma}$ section in Fig. 1a, and further exemplified by KHgSb (in Fig. 3a) with an ideal surface termination. To demonstrate that the surface-localized bands of KHgSb also connect with the surface-resonant bulk bands in this hourglass-flow topology, we modified the surface potential of KHgSb to push the hourglass (along $\widetilde{\Gamma}\widetilde{Z}$) down into the valence band; owing to the proposed hourglass flow, a different hourglass is pulled down from the conduction band along $\widetilde{U}\widetilde{X}$ (see Fig. 3b). In contrast, the second possible connectivity has no robust surface states (see the $\widetilde{X}\widetilde{Z}\widetilde{U}\widetilde{\Gamma}$ section in Fig. 1b).

(3) Along $\widetilde{\Gamma}\widetilde{X}$ ($k_z = 0$), bands divide into two subspaces having either $\overline{M}_z$ eigenvalue $+i$ or $-i$, as follows from $\overline{M}_z^2 = \overline{E}$. As illustrated in Fig. 3b, the two chiral (anti-chiral) surface modes in the $+i$ ($-i$) subspace may be summarized by a mirror Chern number[16]: $\mathcal{C}_{+i} = +2$ ($\mathcal{C}_{-i} = -2$).

Because the 100 surface of KHg*X* also preserves the glideless $\overline{M}_z$, the 100 dispersion along the high-symmetry wavevectors $\overline{\Gamma}\overline{Y}$ in the Brillouin zone (see Fig. 2c) is topologically equivalent to that of the 010 dispersion along $\widetilde{\Gamma}\widetilde{X}$—this reflects two distinct surface projections (illustrated by blue lines in Fig. 2c) of the non-trivial mirror Chern number in the $k_z = 0$ plane (blue plane in Fig. 2c). However, the 100 surface does not respect the glide symmetry ($\overline{M}_x$) that protects the hourglass fermions on the 010 surface. Instead, the 100 surface modes barely disperse with k_z, forming the anisotropic band structure shown in Fig. 4a; the spin expectation value at the Fermi level is shown in

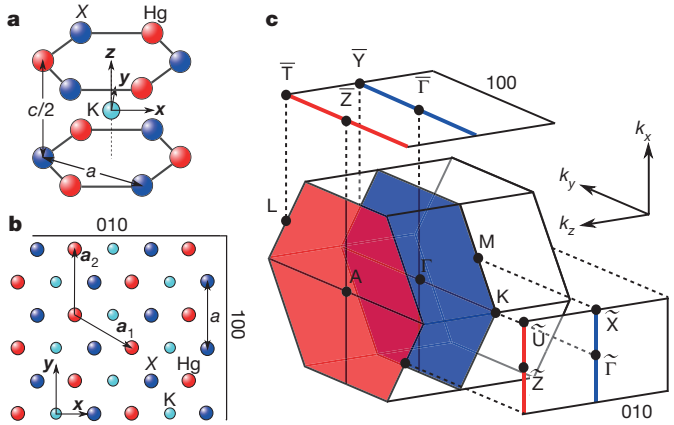


Figure 2 | Crystal structure and Brillouin zone of KHg*X*. a, 3D view of the atomic structure of KHg*X*. The Hg (red) and *X* (blue) ions form a honeycomb layers with AB stacking. The K ion (cyan) is located at an inversion centre, which we choose to be our spatial origin. **b**, Top-down view of a truncated lattice with two surfaces labelled 010 and 100, also known respectively as $(1\bar{2}10)$ and $(10\bar{1}0)$ in the Miller notation. $\mathbf{a}_1$ and $\mathbf{a}_2$ are the planar Bravais lattice vectors. **c**, Centre: bulk Brillouin zone (BZ) of KHg*X*, with two mirror planes of $\overline{M}_z$ coloured red and blue. Top: 100-surface BZ. Right: 010-surface BZ.

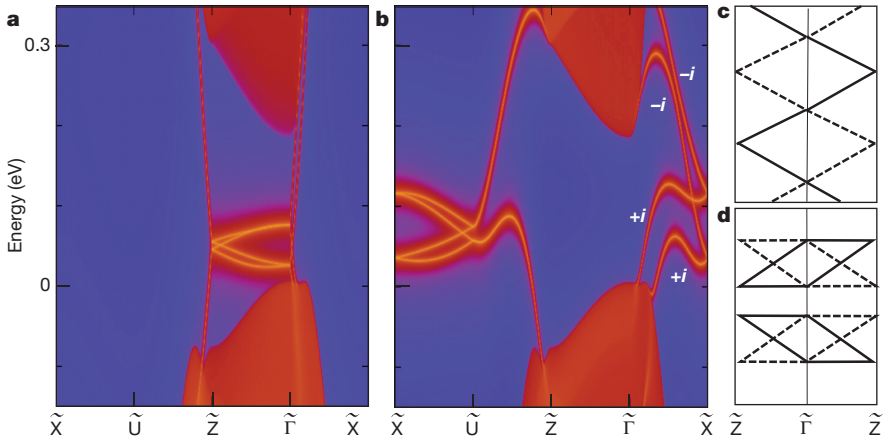


Figure 3 | The 010-surface band structure. a, **b**, The 010-surface bands of KHgSb for an ideal surface termination (**a**) and with a modified surface potential (**b**). Orange lines (the continuum) correspond to the dispersion of surface-localized (bulk-delocalized) electronic states. $\pm i$ labels in **b** refer to the eigenvalues of the mirror operation $\overline{M}_z$ along the mirror line $\widetilde{\Gamma}\widetilde{X}$. **c**, **d**, Possible surface topologies along $\widetilde{Z}\widetilde{\Gamma}\widetilde{Z}$: a glide-symmetric analogue of the quantum spin Hall effect (**c**), and an hourglass connectivity (**d**). Solid and dashed lines distinguish between the two eigenvalue branches of $\overline{M}_x$: $\pm i\exp(-ik_zc/2)$, respectively.

Fig. 4b. The low-energy transport properties are then described by a 3D, doubled QSHE, whereby at each k_z we have two right-moving, spin-down, surface-extended carriers, in combination with their time-reversed partners at $-k_z$.

Bulk analysis

Having enumerated the possible topologies purely from an analysis of the surface symmetries, we proceed to identify which of these topologies are consistent with the bulk symmetries. In a low-energy description of KHgSb, the bulk symmetries are represented by one *s*-type quadruplet[17] (derived from Hg) and three *p*-type quadruplets (from Sb). A 0.2-eV bulk gap is induced by spin–orbit splitting of the *p*-type bands. Supposing electrons fill 12 of these 16 bands, two scenarios emerge. First, if only the *p*-type bands are occupied, as exemplified by KZnP in Fig. 5b, then their corresponding Wannier functions will centre on the P atoms from which the *p* orbitals derive. Second, with KHgSb, the occupied bands along ΓA have mixed *s* and *p* characters (Fig. 5c), which suggests that its Wannier functions centre on the bond between Hg and Sb atoms. Because the 010 surface terminates to produce dangling Hb–Sb bonds (Fig. 2b), the mid-bond Wannier functions of KHgSb mutually hybridize to form surface states. By contrast, the on-atom Wannier functions of KZnP are not expected to form surface states.

Our intuition is justified formally by a Bloch–Wannier representation[18] of the ground state: the n_{occ} occupied bands are represented as hybrid functions $\{|\boldsymbol{k}_\parallel, n\rangle \,|\, n \in \{1, 2, \ldots, n_{\mathrm{occ}}\}\}$ which maximally localize in $\boldsymbol{y}$ (as a Wannier function) but extend in $\boldsymbol{x}$ and $\boldsymbol{z}$ (as a Bloch wave with momentum $\boldsymbol{k}_\parallel = (k_x, k_y)$). Each Bloch–Wannier function is an eigenfunction of the projected-position operator $P_\perp \hat{y} P_\perp$, where $\hat{y}$ is the y-component of the position operator, and $P_\perp$ projects to the occupied bands; the eigenvalue $(y_{n,\boldsymbol{k}_\parallel})$ of $P_\perp \hat{y} P_\perp$ is the centre-of-mass coordinate of the Bloch–Wannier function $(|n, \boldsymbol{k}_\parallel\rangle)$[19]. Owing to the discrete translational symmetry of $P_\perp$, each $y_{n,\boldsymbol{k}_\parallel}$ is a mod-integer quantity representing a family of eigenfunctions related by integer translations; here, we choose the translational period $(a/2)$ in y as the spatial unit $(1 \equiv a/2)$ and a is a lattice constant illustrated in Fig. 2a and b. Because the spectrum of $P_\perp \hat{y} P_\perp$ can be interpolated[20,21] to the 010-surface band structure (Fig. 3a) while preserving the 010 symmetries, we expect[18] that both spectra share similar features (Fig. 6): (i) degenerate doublets along $\widetilde{Z}\widetilde{U}$; (ii) partner-switching quadruplets along $\widetilde{\Gamma}\widetilde{Z}$ and $\widetilde{X}\widetilde{U}$; and (iii) robust crossings between orthogonal $\overline{M}_z$ subspaces (labelled by $\pm i$ in Fig. 6d) along $\widetilde{\Gamma}\widetilde{X}$.

Differences arise because the spectrum of $P_\perp \hat{y} P_\perp$ additionally encodes bulk symmetries that are spoiled by the 010 surface; for example, although our naive surface argument allows for a glide-symmetric QSHE (that is, zigzag connectivity in Fig. 3c) along both $\widetilde{\Gamma}\widetilde{Z}$ and $\widetilde{X}\widetilde{U}$, the out-of-surface translational symmetry rules out this scenario along $\widetilde{X}\widetilde{U}$, as shown in the Supplementary Information. A second difference originates from the bulk inversion $(\mathcal{I})$ symmetry, which quantizes two invariants that have no surface analogue; these invariants describe the polarization of quadruplets along the glide lines $\widetilde{\Gamma}\widetilde{Z}$ and $\widetilde{X}\widetilde{U}$. As an illustration, consider in Fig. 6b the top quadruplet, whose

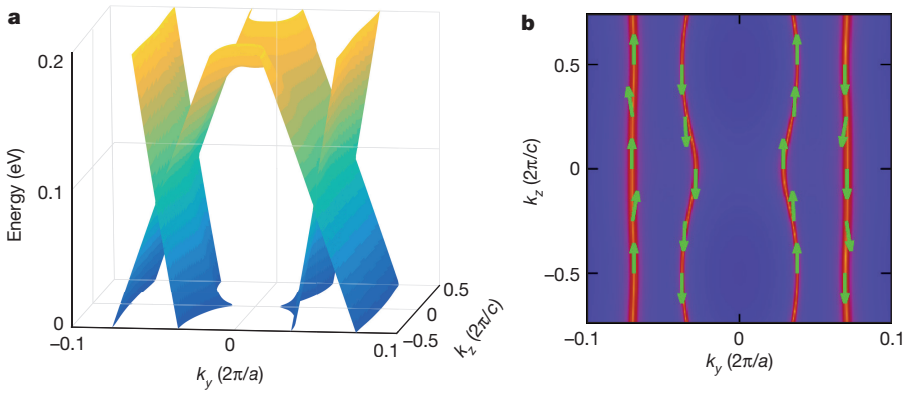


Figure 4 | The 100-surface band structure. a, The 100-surface band structure over a momentum $(\boldsymbol{k}_\parallel)$ rectangle. States with the same colour have the same energy. Although a small hybridization gap (about 1 meV) opens for $k_z \neq 0$, a robust intersection between the $\overline{M}_z = \pm i$ subspaces exists along $k_z = 0$. **b**, Expectation value of the spin (green arrows) for states (orange lines) at the Fermi level.

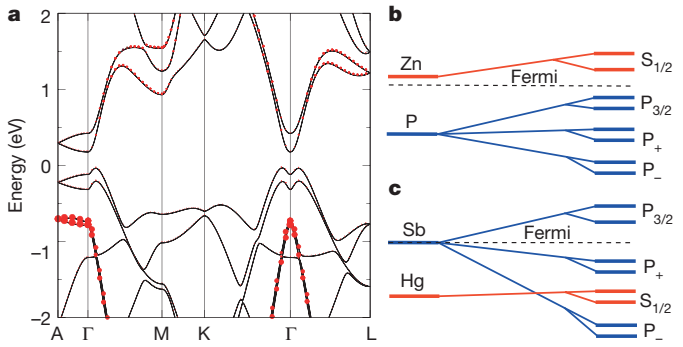


Figure 5 | Bulk band structure and orbital analysis. a, The bulk band structure of KHgSb. The size of each red dot quantifies the weight of the *s* orbitals of Hg. **b**, **c**, Orbital character of KZnP (**b**) and KHgSb (**c**) at any point along ΓA as we vary the crystal field and spin–orbit coupling from zero (left-most) to their natural strengths (right-most). The precise wavefunctions of the electronic states $S_{1/2}$, $P_{3/2}$, P_+ and P_- are clarified in the Supplementary Information. (*s*-type states are shown as orange; *p*-type states are blue.) Unlike KZnP, the ground state of KHgSb has an inverted $S_{1/2}$ quadruplet at all values of the crystal field and spin–orbit coupling, and is consequently metallic where both parameters vanish.

centre-of-mass position ($\mathcal{Y}_1$) may tentatively be defined by averaging four Bloch–Wannier positions: $\mathcal{Y}_1(\boldsymbol{k}_\parallel) = (1/4)\sum_{n=1}^{4} y_{n,\boldsymbol{k}_\parallel}$, with $\boldsymbol{k}_\parallel \in \widetilde{\Gamma}\widetilde{Z}$ and $y_{n,\boldsymbol{k}_\parallel}$ is defined in the previous paragraph and also illustrated in Fig. 6b. Any polarization quantity should be well defined modulo 1, which reflects the discrete translational symmetry of the crystal. However, $\mathcal{Y}$ is only well defined modulo 1/4 for quadruplet bands without symmetry, owing to the integer ambiguity of each of $\{y_n \,|\, n \in \mathbb{Z}\}$. This ambiguity is illustrated in Fig. 6a for an asymmetric insulator with four occupied bands. Only the spectrum for two spatial unit cells (with unit period) is shown, and the discrete translational symmetry ensures $y_{j,\boldsymbol{k}_\parallel} = y_{j+4l,\boldsymbol{k}_\parallel} - l$ for $j, l \in \mathbb{Z}$. Clearly the centres of mass of $\{y_1, y_2, y_3, y_4\}$ and $\{y_2, y_3, y_4, y_5\}$ differ by 1/4 at each $\boldsymbol{k}_\parallel$, but both choices are equally natural given level repulsion across $\widetilde{Z}\widetilde{\Gamma}\widetilde{Z}$. However, a unique choice for the centre of mass exists if the Bloch–Wannier bands divide into sets of four, such that within each set there are enough contact points along $\widetilde{\Gamma}\widetilde{Z}$ to continuously travel between the four bands. Such a property, which we call fourfold connectivity, is illustrated in Fig. 6b for a glide-symmetric insulator with four occupied bands ($n_{\mathrm{occ}} = 4$). Here, both quadruplets $\{y_1, y_2, y_3, y_4\}$ and $\{y_5, y_6, y_7, y_8\}$ are connected, and their centres of mass differ by unity. Our definition of a mod-1 centre-of-mass coordinate then hinges on this fourfold connectivity that characterizes insulators with glide and time-reversal symmetries. To extend this definition to multiple quadruplets per unit cell (where the integer $n_{\mathrm{occ}}/4 \geq 1$), let us define the net displacement of all $n_{\mathrm{occ}}/4$ connected-quadruplet centres as $\mathcal{Q}(\boldsymbol{k}_\parallel)/e = \sum_{j=1}^{n_{\mathrm{occ}}/4} \mathcal{Y}_j(\boldsymbol{k}_\parallel) \bmod 1$, where e is the charge of the electron; this quantity is quantized to either 0 or 1/2 owing to a combination of time-reversal (T) and spatial-inversion ($\mathcal{I}$) symmetry. Indeed, $T\mathcal{I}$ inverts the spatial coordinate but leaves momentum untouched: $T\mathcal{I}|\boldsymbol{k}_\parallel, n\rangle = |\boldsymbol{k}_\parallel, m\rangle$ with $m \neq n$ and $y_{n,\boldsymbol{k}_\parallel} = -y_{m,\boldsymbol{k}_\parallel} \bmod 1$. Consequently, $T\mathcal{I}: \mathcal{Y}_j(\boldsymbol{k}_\parallel) \to \mathcal{Y}_{j'}(\boldsymbol{k}_\parallel) = -\mathcal{Y}_j(\boldsymbol{k}_\parallel) \bmod 1$, and the only non-integer contribution to $\mathcal{Q}/e$ ($\mathcal{Q}/e = 1/2$) arises if there exists a $T\mathcal{I}$-invariant quadruplet ($\bar{j}$) centred at $\mathcal{Y}_{\bar{j}} = 1/2 = -\mathcal{Y}_{\bar{j}} \bmod 1$. Because each $y_{n,\boldsymbol{k}_\parallel}$ is a continuous function of $\boldsymbol{k}_\parallel$, $\mathcal{Q}(\boldsymbol{k}_\parallel)$ is constant ($\mathcal{Q}(\boldsymbol{k}_\parallel) \equiv \mathcal{Q}_{\widetilde{\Gamma}\widetilde{Z}}$) over $\widetilde{\Gamma}\widetilde{Z}$. Alternatively stated, $\mathcal{Q}_{\widetilde{\Gamma}\widetilde{Z}}$ is a quantized polarization invariant that characterizes the entire glide plane that projects to $\widetilde{\Gamma}\widetilde{Z}$. Similarly reasoning with $\widetilde{X}\widetilde{U}$, we obtain two $\mathbb{Z}_2$ invariants: $\mathcal{Q}_{\widetilde{\Gamma}\widetilde{Z}}$ and $\mathcal{Q}_{\widetilde{X}\widetilde{U}}$.

For KHgSb, Fig. 6c illustrates the absence (presence) of the $\mathcal{Y} = 1/2$ quadruplet along $\widetilde{X}\widetilde{U}$ ($\widetilde{\Gamma}\widetilde{Z}$), leading to $\mathcal{Q}_{\widetilde{X}\widetilde{U}} = 0$ and $\mathcal{Q}_{\widetilde{\Gamma}\widetilde{Z}} = e/2$—this difference originates from the band inversion along ΓA (compare with Fig. 5). Wherever $\mathcal{Q}_{\widetilde{\Gamma}\widetilde{Z}} \neq \mathcal{Q}_{\widetilde{X}\widetilde{U}}$, we obtain the hourglass-flow topology exemplified in Fig. 6c. In contrast, Fig. 6g, h depicts the trivial spectrum for KZnP. As initially motivated, $\mathcal{Q}_{\widetilde{\Gamma}\widetilde{Z}} = e/2$ in KHgSb indicates the mid-bond Bloch–Wannier functions, which further hybridize to form the hourglass of Fig. 3a when the 010 surface is terminated.

The topological distinction between KHgSb and KZnP may further be deduced by their differing quantum numbers under spatial transformations. Thus far, the most successful strategy[22] for finding topological materials lies in identifying centrosymmetric systems with inverted parity quantum numbers[16,23–25]. For KHgSb, the parity eigenvalues of the *s* quadruplet (recall Fig. 5) are identical to those of any *p* quadruplet, and therefore there is no parity inversion at any inversion-invariant momentum[26]. Instead, KHgSb manifests an inversion of its eigenvalues ($\exp(-i\pi J_z/3)$) under the screw $\overline{C}_{6z}\cdot[\overline{C}_{6z}, \overline{M}_z] = 0$ implies states at Γ can simultaneously be labelled by both operators. The $\overline{M}_z = +i$ states in the *s*-quadruplet (*p*-quadruplet) transform as $J_z = -1/2$ and $J_z = 5/2$ ($J_z = 3/2$ and $J_z = -3/2$), and the inversion

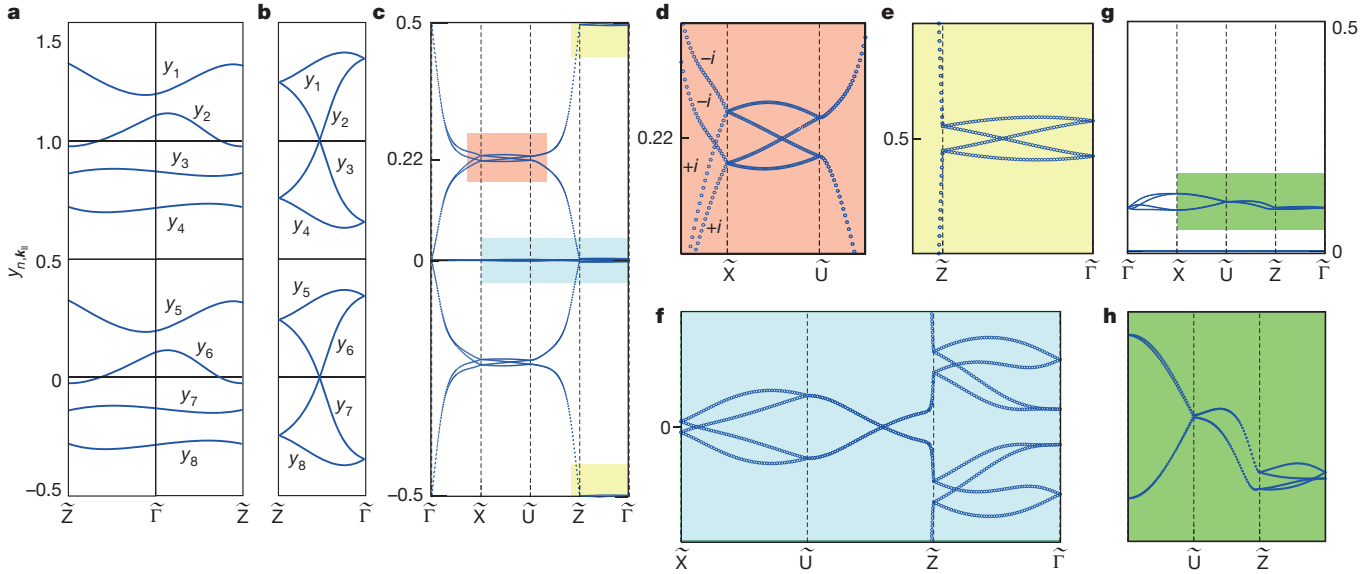


Figure 6 | Spectra of the projected-position operator $P_\perp \hat{\boldsymbol{y}} P_\perp$. **a**, **b**, Comparison of the spectrum of $P_\perp \hat{\boldsymbol{y}} P_\perp$ for a system without any symmetry (**a**) and one with time-reversal, spatial-inversion and glide symmetries (**b**). The vertical axis shows position in units of the translational period in the $\boldsymbol{y}$ direction. **c**, Spectrum of $P_\perp \hat{\boldsymbol{y}} P_\perp$ for KHgSb. **d**–**f**, Close-ups of the pink (**d**), light green (**e**) and blue (**f**) regions of the spectrum in **c**. **g**, Spectrum of $P_\perp \hat{\boldsymbol{y}} P_\perp$ for KZnP in half the unit cell. **h**, Close-up of the dark green region of the spectrum in **g**.

between these s- and p-states at Γ results in a net angular momentum gain of $\Delta J_z = 2$, which accompanies a quantized redistribution of Berry curvature[27]; that is, ΔJ_z is equal to the change in $\mathcal{C}_{+i}$ modulo 6, as proven in the Supplementary Information. There, we further confirm $\mathcal{C}_{+i} = 2$ using the Wilson-loop method, in accordance with our surface analysis.

Discussion

Spatial symmetries have played a crucial role in the topological classification of band insulators[23,28–34]; non-symmorphic spatial symmetries are particularly useful in the classification of band semimetals[35,36] and their Fermi-liquid analogues[37], and in the identification of topologically ordered insulators with fractionalized excitations[38–40]. So far, all experimentally tested topological insulators have relied on symmorphic space groups[2,23–25]. In KHgX, we propose a family of insulators with non-symmorphic topology, in the hope of stimulating interest in an experimentally barren field. Our time-reversal-invariant theory of KHgX complements previous theoretical proposals with magnetic, non-symmorphic space groups[29,31,41–43]. We propose to characterize glide-symmetric crystals, such as KHgX, by a quantized polarization that depends on the non-Abelian Berry connection[19,44,45]. In constrast, the standard polarization relates to the Abelian Berry connection[9,12,27]. Additionally, KHgX uniquely exemplifies a 'rotationally inverted' insulator; a general strategy to search for such materials in all space groups is elaborated in the Supplementary Information.

KHgX represents one of many possible topologies within its space group, as illustrated in Fig. 1a–c. To determine which symmetry groups, other than that of KHgX, allow for topological surface states, we propose a criterion on the surface symmetry that applies to all known symmetry protected surface topologies. By 'symmetry protected', we mean non-chiral surface states with vanishing Chern[46] (or mirror Chern[16]) numbers. Our criterion introduces the notion of connectivity within a submanifold ($\mathcal{M}$) of the surface Brillouin zone, and relates to the theory of elementary energy bands[35,47]—we say that $\mathcal{M}$ is $\mathcal{D}$-fold connected if bands there divide into sets of $\mathcal{D}$, such that within each set there are enough contact points in $\mathcal{M}$ to continuously travel through all $\mathcal{D}$ bands. If $\mathcal{M}$ is a single wavevector ($\boldsymbol{k}_\parallel$), then $\mathcal{D}$ coincides with the dimension of the irreducible representation at $\boldsymbol{k}_\parallel$; $\mathcal{D}$ generalizes this notion of symmetry enforced degeneracy where $\mathcal{M}$ is larger than a wavevector (for example, a glide line). Our criterion is as follows: (a) there exist two separated submanifolds $\mathcal{M}_1$ and $\mathcal{M}_2$, with corresponding $\mathcal{D}_1 = \mathcal{D}_2 = fd$ ($f \geq 2$ and $d \geq 1$ are integers); and (b) there exists a third submanifold $\mathcal{M}_3$ that connects $\mathcal{M}_1$ and $\mathcal{M}_2$, with corresponding $\mathcal{D}_3 = d$. Almost all symmetry protected surface topologies[3,28–30,48] are characterized by $\mathcal{D}_1 = \mathcal{D}_2 = 2\mathcal{D}_3 = 2$, with $\mathcal{M}_1$ and $\mathcal{M}_2$ being two high-symmetry wavevectors connected by a curve $\mathcal{M}_3$; for example, the edge of the quantum-spin-Hall insulator[15] is characterized by two Kramers-degenerate momenta (hence $\mathcal{D}_1 = \mathcal{D}_2 = 2$) connected by a curve with trivial degeneracy ($\mathcal{D}_3 = 1$)—these constraints allow for a Kramers-partner-switching dispersion[3]. Here, the surface symmetry $Pma2$ is characterized by two glide lines ($\mathcal{M}_1 = \widetilde{\Gamma}\widetilde{\mathrm{Z}}$ and $\mathcal{M}_2 = \widetilde{\mathrm{XU}}$) with hourglass band structures ($\mathcal{D}_1 = \mathcal{D}_2 = 4$), and a glideless mirror line ($\mathcal{M}_3 = \widetilde{\mathrm{ZU}}$) with doubly-degenerate bands ($\mathcal{D}_3 = 2$). Previous studies of magnetic systems[34,42,43,49] have established a $\mathbb{Z}_2$ topology with $\mathcal{D}_1 = \mathcal{D}_2 = 2\mathcal{D}_3 = 2$, where $\mathcal{M}_1$ and $\mathcal{M}_2$ are also parallel glide lines. Our surface-centric criterion for non-trivial topology is sometimes over-predictive because it neglects bulk symmetries that are spoiled by the surface—a fully predictive methodology involves the representation theory of Wilson loops and the notion of a cohomological insulator[17]. Finally, an exciting direction for future research lies in gapping the hourglass fermion with magnetism and superconductivity[50].

Received 1 November 2015; accepted 4 February 2016.

1. Lax, M. *Symmetry Principles in Solid State and Molecular Physics* (Ch. 8 Wiley-Interscience, 1974).
2. Hsieh, D. *et al.* A topological Dirac insulator in a quantum spin Hall phase. *Nature* **452,** 970–974 (2008).
3. Kane, C. L. & Mele, E. J. Z_2 topological order and the quantum spin Hall effect. *Phys. Rev. Lett.* **95,** 146802 (2005).
4. Bernevig, B. A., Hughes, T. L. & Zhang, S. C. Quantum spin Hall effect and topological phase transition in HgTe quantum wells. *Science* **314,** 1757–1761 (2006).
5. König, M. *et al.* Quantum spin Hall insulator state in HgTe quantum wells. *Science* **318,** 766–770 (2007).
6. Fu, L., Kane, C. L. & Mele, E. J. Topological insulators in three dimensions. *Phys. Rev. Lett.* **98,** 106803 (2007).
7. Moore, J. E. & Balents, L. Topological invariants of time-reversal-invariant band structures. *Phys. Rev. B* **75,** 121306 (2007).
8. Roy, R. Z_2 classification of quantum spin Hall systems: an approach using time-reversal invariance. *Phys. Rev. B* **79,** 195321 (2009).
9. King-Smith, R. D. & Vanderbilt, D. Theory of polarization of crystalline solids. *Phys. Rev. B* **47,** 1651–1654 (1993).
10. Vanderbilt, D. & King-Smith, R. D. Electric polarization as a bulk quantity and its relation to surface charge. *Phys. Rev. B* **48,** 4442–4455 (1993).
11. Resta, R. Macroscopic polarization in crystalline dielectrics: the geometric phase approach. *Rev. Mod. Phys.* **66,** 899–915 (1994).
12. Zak, J. Berry's phase for energy bands in solids. *Phys. Rev. Lett.* **62,** 2747–2750 (1989).
13. Vogel, R. & Schuster, H.-U. KHgAs (Sb) und KZnAs - ternäre Verbindungen mit modifizierter Ni_2In-Struktur. *Z. Naturforsch.* **35b,** 114–116 (1980).
14. Tinkham, M. *Group Theory and Quantum Mechanics* 279–281 (Dover, 2003).
15. Kane, C. L. & Mele, E. J. Quantum spin Hall effect in graphene. *Phys. Rev. Lett.* **95,** 226801 (2005).
16. Teo, J. C. Y., Fu, L. & Kane, C. L. Surface states and topological invariants in three-dimensional topological insulators: application to $Bi_{1-x}Sb_x$. *Phys. Rev. B* **78,** 045426 (2008).
17. Alexandradinata, A., Wang, Z. & Bernevig, B. A. Topological insulators by group cohomology. *Phys. Rev. X* (in the press).
18. Taherinejad, M., Garrity, K. F. & Vanderbilt, D. Wannier center sheets in topological insulators. *Phys. Rev. B* **89,** 115102 (2014).
19. Alexandradinata, A., Dai, X. & Bernevig, B. A. Wilson-loop characterization of inversion-symmetric topological insulators. *Phys. Rev. B* **89,** 155114 (2014).
20. Fidkowski, L., Jackson, T. S. & Klich, I. Model characterization of gapless edge modes of topological insulators using intermediate Brillouin-zone functions. *Phys. Rev. Lett.* **107,** 036601 (2011).
21. Huang, Z. & Arovas, D. P. Entanglement spectrum and Wannier center flow of the Hofstadter problem. *Phys. Rev. B* **86,** 245109 (2012).
22. Fu, L. & Kane, C. L. Topological insulators with inversion symmetry. *Phys. Rev. B* **76,** 045302 (2007).
23. Hsieh, T. H. *et al.* Topological crystalline insulators in the SnTe material class. *Nature Commun.* **3,** 982 (2012).
24. Xu, S.-Y. *et al.* Observation of a topological crystalline insulator phase and topological phase transition in $Pb_{1-x}Sn_xTe$. *Nature Commun.* **3,** 1192 (2012).
25. Tanaka, Y. *et al.* Experimental realization of a topological crystalline insulator in SnTe. *Nature Phys.* **8,** 800–803 (2012).
26. Zhang, H.-J. *et al.* Topological insulators in ternary compounds with a honeycomb lattice. *Phys. Rev. Lett.* **106,** 156402 (2011).
27. Berry, M. V. Quantal phase factors accompanying adiabatic changes. *Proc. R. Soc. Lond. A* **392,** 45–57 (1984).
28. Fu, L. Topological crystalline insulators. *Phys. Rev. Lett.* **106,** 106802 (2011).
29. Liu, C. X., Zhang, R. X. & VanLeeuwen, B. K. Topological nonsymmorphic crystalline insulators. *Phys. Rev. B* **90,** 085304 (2014).
30. Alexandradinata, A., Fang, C., Gilbert, M. J. & Bernevig, B. A. Spin-orbit-free topological insulators without time-reversal symmetry. *Phys. Rev. Lett.* **113,** 116403 (2014).
31. Fang, C., Gilbert, M. J. & Bernevig, B. A. Entanglement spectrum classification of C_n-invariant noninteracting topological insulators in two dimensions. *Phys. Rev. B* **87,** 035119 (2013).
32. Shiozaki, K. & Sato, M. Topology of crystalline insulators and superconductors. *Phys. Rev. B* **90,** 165114 (2014).
33. Po, H. C., Watanabe, H., Zaletel, M. P. & Vishwanath, A. Filling-enforced quantum band insulators in spin-orbit coupled crystals. Preprint at http://arxiv.org/abs/1506.03816 (2015).
34. Varjas, D. *et al.* Bulk invariants and topological response in insulators and superconductors with nonsymmorphic symmetries. *Phys. Rev. B* **92,** 195116 (2015).
35. Michel, L. & Zak, J. Elementary energy bands in crystalline solids. *Europhys. Lett.* **50,** 519–525 (2000).
36. Young, S. M. & Kane, C. L. Dirac semimetals in two dimensions. *Phys. Rev. Lett.* **115,** 126803 (2015).
37. Parameswaran, S. A. Topological 'Luttinger' invariants protected by crystal symmetry in semimetals. Preprint at http://arxiv.org/abs/1508.01546 (2015).
38. Parameswaran, S. A. *et al.* Topological order and absence of band insulators at integer filling in non-symmorphic crystals. *Nature Phys.* **9,** 299–303 (2013).
39. Roy, R. Space group symmetries and low lying excitations of many-body systems at integer fillings. Preprint at http://arxiv.org/abs/1212.2944 (2012).
40. Watanabe, H. *et al.* Filling constraints for spin-orbit coupled insulators in symmorphic and nonsymmorphic crystals. *Proc. Natl Acad. Sci. USA* **112,** 14551–14556 (2015).
41. Mong, R. S. K., Essin, A. M. & Moore, J. E. Antiferromagnetic topological insulators. *Phys. Rev. B* **81,** 245209 (2010).

42. Fang, C. & Fu, L. New classes of three-dimensional topological crystalline insulators: nonsymmorphic and magnetic. *Phys. Rev. B* **91,** 161105 (2015).
43. Shiozaki, K., Sato, M. & Gomi, K. Z_2 topology in nonsymmorphic crystalline insulators: Möbius twist in surface states. *Phys. Rev. B* **91,** 155120 (2015).
44. Wilczek, F. & Zee, A. Appearance of gauge structure in simple dynamical systems. *Phys. Rev. Lett.* **52,** 2111–2114 (1984).
45. Alexandradinata, A. & Bernevig, B. A. Berry-phase description of topological crystalline insulators. Preprint at http://arxiv.org/abs/1409.3236 (2014).
46. Haldane, F. D. M. Model for a quantum Hall effect without Landau levels: condensed-matter realization of the "parity anomaly". *Phys. Rev. Lett.* **61,** 2015–2018 (1988).
47. Michel, L. & Zak, J. Connectivity of energy bands in crystals. *Phys. Rev. B* **59,** 5998–6001 (1999).
48. Dong, X.-Y. & Liu, C.-X. The classification of topological crystalline insulators based on representation theory. *Phys. Rev. B* **93,** 045429 (2016).
49. Lu, L. *et al.* Symmetry-protected topological photonic crystal in three dimensions. *Nature Phys*. http://dx.doi.org/10.1038/nphys3611 (2016).
50. Fu, L. & Kane, C. L. Superconducting proximity effect and Majorana fermions at the surface of a topological insulator. *Phys. Rev. Lett.* **100,** 096407 (2008).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank C. Fang, D. P. Arovas, J. Li, L. Muechler and X. Dai for discussions. This work was supported by NSF CAREER DMR-095242, ONR-N00014-11-1-0635, TI MURI W911NF-12-1-0461, NSF-MRSEC DMR-1420541, Packard Foundation, Keck grant, "ONR Majorana Fermions" 25812-G0001-10006242-101, and Schmidt fund 23800-E2359-FB625. During the refereeing stages of this work, A.A. was supported by the Yale Fellowship in Condensed Matter Physics.

Author Contributions A.A., Z.W. and B.A.B. performed theoretical analysis; Z.W. discovered the KHg*X* material class and performed the first-principles calculations; R.J.C. provided several other material suggestions.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.A.B. (bernevig@princeton.edu).

ARTICLE

doi:10.1038/nature17623

Astrocyte scar formation aids central nervous system axon regeneration

Mark A. Anderson[1]*†, Joshua E. Burda[1]*, Yilong Ren[1]†, Yan Ao[1], Timothy M. O'Shea[1], Riki Kawaguchi[2], Giovanni Coppola[2], Baljit S. Khakh[3], Timothy J. Deming[4] & Michael V. Sofroniew[1]

Transected axons fail to regrow in the mature central nervous system. Astrocytic scars are widely regarded as causal in this failure. Here, using three genetically targeted loss-of-function manipulations in adult mice, we show that preventing astrocyte scar formation, attenuating scar-forming astrocytes, or ablating chronic astrocytic scars all failed to result in spontaneous regrowth of transected corticospinal, sensory or serotonergic axons through severe spinal cord injury (SCI) lesions. By contrast, sustained local delivery via hydrogel depots of required axon-specific growth factors not present in SCI lesions, plus growth-activating priming injuries, stimulated robust, laminin-dependent sensory axon regrowth past scar-forming astrocytes and inhibitory molecules in SCI lesions. Preventing astrocytic scar formation significantly reduced this stimulated axon regrowth. RNA sequencing revealed that astrocytes and non-astrocyte cells in SCI lesions express multiple axon-growth-supporting molecules. Our findings show that contrary to the prevailing dogma, astrocyte scar formation aids rather than prevents central nervous system axon regeneration.

Transected axons fail to regrow spontaneously across severe tissue lesions in the mature mammalian central nervous system (CNS). Potential mechanisms include: (i) reduced intrinsic growth capacity of mature CNS neurons[1–3]; (ii) absence of external growth stimulating and supporting factors[1,4,5]; and (iii) presence of external inhibitory factors associated with myelin[6,7], fibrotic tissue[8] or astrocytic scars[9]. Alleviating cellular and molecular mechanisms underlying axon regeneration failure is fundamental to improving CNS repair after traumatic injury, stroke or degenerative disease.

Astrocytic scars have been regarded as barriers to CNS axon regrowth since the mid-twentieth century on the basis of their appearance and early reports that attenuating astrocytic scar formation enabled spontaneous axon regrowth[10,11]. Although axon-growth-promoting effects of early scar attenuators proved illusory, reports correlating failed axon regrowth with presence of mature astrocytes[12] or astrocytic scars[9], plus evidence that astrocytes produce chondroitin sulfate proteoglycans (CSPGs) that inhibit axon growth *in vitro*[9], led to the widespread view that astrocytic scars are critical inhibitors of CNS axons and that nullifying this inhibition will lead to spontaneous axon regeneration.

Here we tested the hypothesis that astrocytic scar formation plays a causal role in the failure of transected mature CNS axons to regenerate across severe tissue lesions. We used multiple transgenic loss-of-function strategies to ablate scar-forming astrocytes, genetically attenuate scar-forming astrocytes or to ablate chronic astrocytic scars after severe SCI in adult mice. We quantified the effects of these manipulations on (i) spontaneous regeneration of three major types of CNS axons; (ii) total CSPG levels; (iii) genome-wide expression by astrocytes and non-astrocytes in SCI lesions of molecules associated with axon growth; and (iv) axon regeneration stimulated by conditioning lesions plus delivery via synthetic hydrogel depots of known axon-required growth factors missing from SCI lesions.

No axon regrowth after preventing scars

We first determined effects of preventing astrocytic scar formation on the potential for spontaneous unstimulated axon regeneration through severe CNS lesions. After focal traumatic tissue damage, CNS lesions comprise central areas of non-neural lesion core tissue surrounded by narrow astrocytic scar borders[13,14]. Astrocytic scar formation is complete by two weeks after adult murine SCI and is critically dependent on astrocyte proliferation and STAT3 signalling[15–18]. We prevented astrocyte scar formation with two loss-of-function transgenic mouse models that either selectively kill proliferating scar-forming astrocytes[15,16] or delete STAT3 signalling selectively from astrocytes[17,18], referred to respectively as TK+GCV or STAT3-CKO mice (Supplementary Information).

After severe crush SCI, wild-type mice formed dense astrocytic scars by two weeks that persisted for eight weeks, whereas TK+GCV and STAT3-CKO mice failed to form scars and instead exhibited larger areas of non-neural tissue around lesion centres that were essentially devoid of astrocytes from two to eight weeks after SCI (Fig. 1a–d, Extended Data Fig. 1 and Supplementary Information).

Effects of preventing astrocyte scar formation on axon regeneration were quantified in three axonal systems: (i) descending corticospinal tract (CST); (ii) ascending sensory tract (AST); and (iii) descending serotonergic (5HT) tract, visualized either by axonal tract tracing or immunohistochemistry (Fig. 1c, e–i and Supplementary Information). As expected after severe SCI in adult wild-type mice, transected CST and AST axons both exhibited moderate dieback away from lesion centres (Fig. 1c, e, g, h). Preventing astrocytic scar formation in either TK+GCV or STAT3-CKO mice failed to result in spontaneous regrowth of transected CST or AST axons through SCI lesions, and instead significantly increased axonal dieback (Fig. 1c, e, g, h). As expected[19,20], transected 5HT axons exhibited little dieback from lesion centres (Fig. 1f, i). Preventing scar formation in TK+GCV or STAT3-CKO mice did not exacerbate dieback of 5HT axons.

[1]Department of Neurobiology, David Geffen School of Medicine, University of California, Los Angeles, California 90095-1763, USA. [2]Departments of Psychiatry and Neurology, David Geffen School of Medicine, University of California, Los Angeles, California 90095-1761, USA. [3]Department of Physiology, David Geffen School of Medicine, University of California, Los Angeles, California 90095-1751, USA. [4]Departments of Bioengineering, Chemistry and Biochemistry, University of California Los Angeles, Los Angeles, California 90095-1600, USA. †Present addresses: School of Life Sciences, Swiss Federal Institute of Technology (EPFL), SV BMI UPCourtine, Station 19, CH-1015 Lausanne, Switzerland (M.A.A); Department of Spine Surgery, Tongji Hospital, Tongji University School of Medicine, Shanghai 200065, China (Y.R.).
*These authors contributed equally to this work.

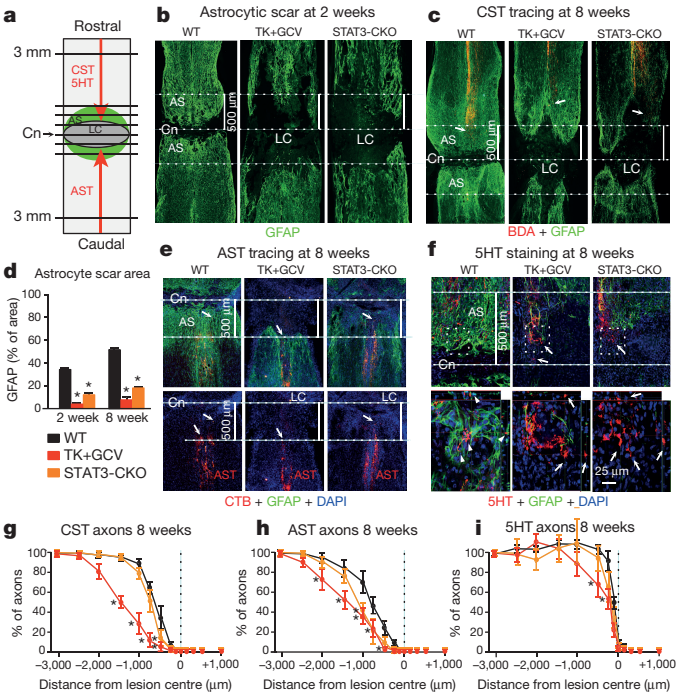


Figure 1 | Preventing astrocytic scar formation does not lead to spontaneous regrowth of CST, AST or 5HT axons after SCI. **a**, Experiment summary schematic. Horizontal view of lesion core (LC) and astrocytic scar (AS) after SCI. Intercepts of CST, AST or 5HT axons with lines drawn at various distances from lesion centre (Cn) were counted and expressed as a percentage of all axons 3 mm proximal. **b**, **c**, **e**, **f**, Dotted lines demarcate lesion centre and 500 μm on either side. **b**, **d**, Area occupied by GFAP-positive scar-forming astrocytes within 500 μm either side of the lesion centre at 2 or 8 weeks after SCI. $n = 6$. **c**, **e**, Arrows depict most-caudal CST axons (biotinylated dextran amine tracing) or most-rostral AST axons (choleratoxin B tracing). **f**, Arrows in top images depict most caudally penetrating 5HT axons, boxed areas are shown below. In wild-type mice, 5HT axons are surrounded by the astrocytic scar (arrowheads). In TK+GCV and STAT3-CKO mice, many 5HT axons are not in contact with the astrocytic scar (arrows), but have not regrown. **g**–**i**, Numbers (means ± s.e.m.) of CST (**g**), AST (**h**) and 5HT (**i**) axons at various distances from the SCI lesion centre as a percentage of the total number of axons present 3 mm proximal. CST, $n = 6$ mice; AST and 5HT, $n = 5$ mice; $^*P < 0.05$ versus wild type (ANOVA with Newman–Keuls).

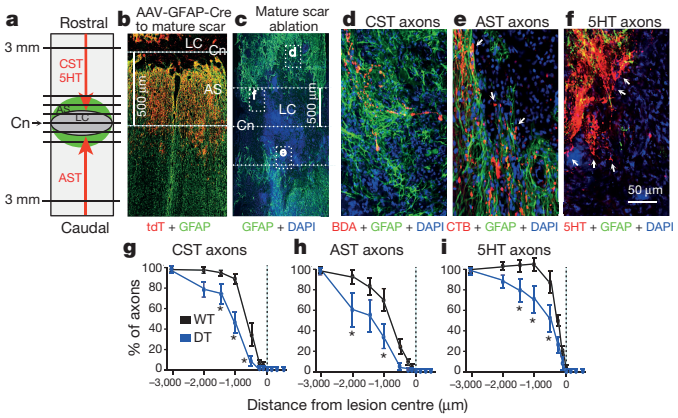


Figure 2 | No spontaneous regrowth of CST, AST or 5HT axons after ablating chronic astrocyte scars. **a**, Experiment summary schematic. **b**, Selective targeting of tdTomato reporter to GFAP-positive scar-forming astrocytes in 500 μm zone occupied by the astrocytic scar (see Extended Data Fig. 1c). **c**, Diphtheria toxin-mediated ablation of chronic astrocytic scar after severe SCI in mice with transgenically targeted diphtheria toxin receptor. Dotted lines indicate lesion centre (Cn) and 500 μm on either side normally occupied by scar-forming astrocytes in wild-type (WT) mice. Boxes show locations of **d**–**f** in adjacent sections. **d**, CST axons are found only among GFAP-positive astrocytes proximal to ablated scar. **e**, AST axons are present at the margins of a large area depleted of scar but have not regrown (arrows). **f**, 5HT axons are present within an area depleted of scar but have not regrown (arrows). **g**–**i**, Numbers (means ± s.e.m.) of CST (**g**), AST (**h**), or 5HT (**i**) axons at various distances from SCI lesion centres as a percentage of the total number of axons present at 3 mm proximal. CST, $n = 6$ mice; AST and 5HT, $n = 5$ mice; $^*P < 0.05$ versus wild type (ANOVA with Newman–Keuls).

Nevertheless, although many 5HT axons remained in lesion centres devoid of astrocytes, they also failed to regrow (Fig. 1f, i).

Thus, in spite of the essential absence of scar-forming astrocytes from SCI lesions for eight weeks after SCI in TK + GCV mice or STAT3-CKO mice, there was no spontaneous regeneration of transected CST, AST or 5HT axons through the lesions. This regrowth failure was particularly apparent for AST and 5HT axons, the axonal tips of which were often present along or within large areas devoid of astrocytes but did not regrow spontaneously through such areas (Fig. 1e, f).

No axon regrowth after removing chronic scars

Acute astrocytic scar formation restricts inflammation and preserves neural tissue[14–18]. It has been proposed that after inflammation has resolved, chronic astrocytic scars are expendable and detrimental because they continually prevent axon regeneration. To test this hypothesis, we ablated chronic astrocytic scars five weeks after SCI with genetically targeted diphtheria toxin receptor and ultra-low doses of diphtheria toxin[21] (Fig. 2 and Supplementary Information). Distribution and specificity of targeting to mature astrocytic scars was verified with the genetic reporter tdTomato (Fig. 2b, Extended Data Fig. 1c and Supplementary Information). GFAP immunohistochemistry verified efficient removal of chronic astrocytic scars (Fig. 2c). Axon quantitation ten weeks after SCI showed that transected CST, AST or 5HT axons all failed to regrow spontaneously through areas depleted of chronic astrocytic scars (Fig. 2d–i). Again, this failure was particularly notable for AST and 5HT axons in or along areas devoid of scar-forming astrocytes that did not regrow through these areas (Fig. 2e, f). We also ablated chronic astrocytic scars and adjacent astrocytes over larger areas to reach 'died-back' CST and AST axons, but this approach caused pronounced tissue degeneration and large lesions (Extended Data Fig. 1e) that contained essentially no detectable CST, AST or 5HT axons. These findings show that ablating chronic astrocytic scars fails to result in spontaneous regrowth of CST, AST or 5HT axons through SCI lesions, and that chronic astrocytic scars remain critical for sustaining tissue integrity.

Multicellular CSPG production

We next looked for molecular mechanisms that might explain why ablating or attenuating astrocytic scars failed to enable spontaneous axon regrowth through severe lesions. CSPGs produced by astrocytic scars are regarded as the principal inhibitors of axon regeneration[9]. Total CSPG levels determined by dot blot with CS56 antibody[9,22] were, as expected[9], significantly higher in our wild-type SCI lesions, but were not significantly reduced by transgenic ablation or disruption of astrocytic scar formation (Fig. 3a). Because diverse cells in SCI lesions including pericytes, fibroblast lineage cells and inflammatory cells[13] can produce CSPGs[23], we examined cellular production of CSPG and GFAP and quantified immunohistochemically stained tissue areas. In SCI lesions of TK+GCV and STAT3-CKO mice, GFAP area was significantly reduced in both grey and white matter compared with wild type, whereas CSPG area was not significantly reduced in lesion core tissue or in regions of ablated astrocytic scars, which were filled with CSPG-positive, GFAP-negative cells (Figs 3b,c, Extended Data Fig. 2 and Supplementary Information). These findings show that non-astrocyte cells in SCI lesions produce substantive amounts of CSPGs

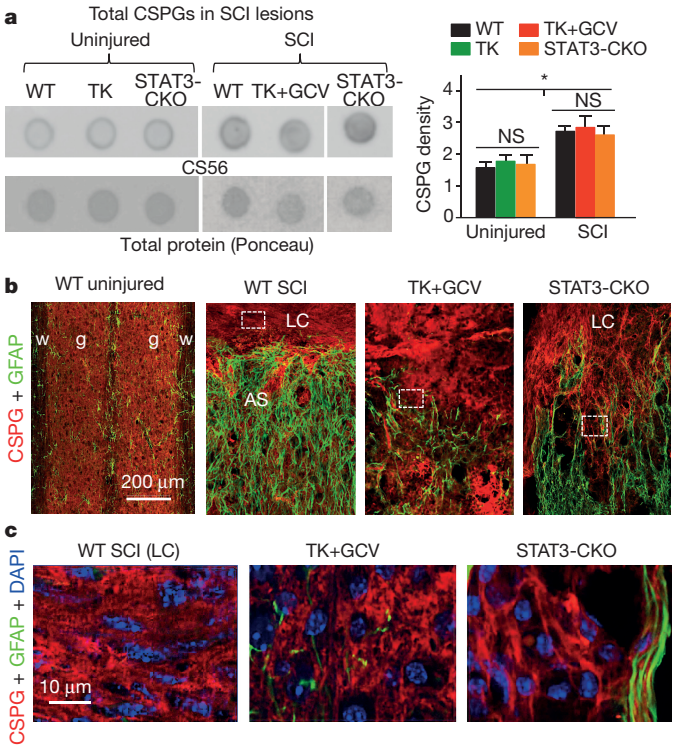


Figure 3 | CSPG production by non-astrocyte cells in SCI lesions after ablation or attenuation of astrocytic scars. a, Dot blots (left) and quantification (right, mean and s.e.m.) of total CSPGs detected by CS56 antibody relative to total protein levels (Ponceau). $n = 4$ mice; $^*P < 0.05$ versus uninjured (ANOVA with Newman–Keuls). **b**, CS56 and GFAP immunohistochemistry (see Extended Data Fig. 2). AS, astrocytic scar; LC, lesion core; g, grey matter; w, white matter. **c**, Details of boxed areas in **b** showing CSPG production by GFAP-negative cells.

and that preventing astrocytic scar formation fails to reduce total CSPG production in SCI lesions.

Genomic dissection of SCI lesions

Numerous molecules attract, repel, support or inhibit axon growth during development, and many of these are present in CNS lesions[7,24,25]. To look broadly at molecules produced by astrocytes or non-astrocyte cells in SCI lesions that might affect axon regrowth, we conducted genome-wide RNA sequencing of astrocyte-specific ribosome-associated RNA (ramRNA) precipitated via a haemagglutinin tag[26] transgenically targeted to either wild-type or STAT3-CKO astrocytes, and non-precipitated (flow-through) RNA deriving from non-astrocyte cells in the same tissue samples (Fig. 4, Extended Data Fig. 3 and Supplementary Information).

At two weeks after SCI, astrocytes and non-astrocyte cells in SCI lesions exhibited significantly altered expression of many genes in both wild-type and STAT3-CKO mice, and wild-type astrocytes exhibited expected known changes[27] (Extended Data Fig. 4a–d and Supplementary Information). Notably, 63% of genes significantly regulated by wild-type astrocytes were not significantly altered by STAT3-CKO astrocytes after SCI, and STAT3-CKO SCI astrocyte transcriptomes clustered more similarly towards uninjured astrocytes than towards wild-type SCI astrocytes (Fig. 4c, d).

We analysed 59 molecules reported to negatively or positively modulate axon growth in SCI lesions (Extended Data Table 1). In SCI lesions from wild-type mice, both astrocytes and non-astrocyte cells expressed a majority not only of 28 known axon inhibitors, including specific CSPGs, ephrins, netrins, neuropillins, plexins, slits and others, but also a majority of 31 known axon permissive molecules, including specific CSPGs, laminins, syndecans, glypicans, decorin and others (Fig. 4e). Notably, both astrocytes and non-astrocytes downregulated more axon inhibitory molecules than were upregulated, and upregulated more than three times the number of axon-permissive molecules than were downregulated. Preventing astrocytic scar formation in STAT3-CKO mice did not significantly decrease the expression by astrocytes or non-astrocytes of a single reported inhibitor, whereas eleven were upregulated, and significantly increased the expression of two permissive molecules, and decreased three, compared to wild-type SCI (Fig. 4e).

In agreement with our immunoblot and immunohistochemical CSPG findings (above), various individual CSPG ramRNAs were expressed by both astrocytes and non-astrocyte cells in SCI lesions (Fig. 4e). Interestingly, aggrecan, the prototypical CSPG used in axon growth inhibition studies *in vitro*[9,28], was not detectably expressed by scar-forming astrocytes at either the ramRNA or immunohistochemistry of protein levels (Fig. 4e and Extended Data Fig. 5a). Other axon-inhibitory CSPGs, brevican, neurocan, versican and phosphacan, were all expressed by both scar-forming astrocytes and non-astrocyte cells in SCI lesions as revealed by RNA analysis (Fig. 4e) and confirmed by immunohistochemistry of protein for brevican and neurocan (Extended Data Figs 5b and 6a).

CSPGs are diverse with respect to inhibiting or supporting axon growth at both protein and sugar-epitope levels[29,30]. Of the five growth-inhibitory CPSGs, wild-type scar-forming astrocytes increased above normal only versican ramRNA, and significantly decreased neurocan and phosphacan (Fig. 4e). In contrast, ramRNAs of two axon-growth-supportive CSPGs, *Cspg4* (NG2) and *Cspg5* (neuroglycan C) (Extended Data Table 1), were significantly upregulated by scar-forming astrocytes (Fig. 4e) and both NG2 and CSPG5 clearly decorated scar-forming astrocytes, as revealed by immunohistochemistry (Extended Data Fig. 6b, c).

Our genomic findings show that: (i) STAT3-CKO prevents or attenuates a majority of genome-wide changes in astrocytes associated with astrogliosis and scar formation in wild-type mice; (ii) astrocytes and non-astrocyte cells in SCI lesions express a large, diverse mix of axon inhibitory and permissive molecules; (iii) non-astrocyte cells in SCI lesions substantively express CSPGs; (iv) preventing astrocyte scar formation with STAT3-CKO does not reduce expression of CSPGs or other inhibitory molecules in SCI lesions; (v) scar-forming astrocytes upregulate and substantively express axon-growth-supporting CSPGs, indicating that CS56 immune detection of total CSPG levels[22] need not indicate a purely axon-inhibitory environment; and (vi) scar-forming astrocytes and non-astrocyte cells in SCI lesions upregulate multiple axon-growth-permissive matrix molecules, including laminins.

Axon regrowth in spite of scar formation

We next stimulated axon growth after SCI in the presence or absence of astrocytic scar formation. Developing axons do not grow by default but require stimulatory cues[31]. This requirement may apply also to regrowth of transected mature axons. Some transected mature AST axons can be stimulated to regrow in severe SCI lesions by activating neuron intrinsic growth programs with peripheral conditioning lesions[32–34], and this regrowth can be significantly augmented by cell grafts that provide supportive matrix plus the neurotrophic factors NT3 and BDNF that attract AST axon growth during development[35]. We noted the essential absence of *Nt3* and *Bdnf* expression in our wild-type SCI lesions, combined with expression of permissive matrix molecules including laminins known to support developing AST axons[36] (Fig. 4e and Extended Data Table 1). We therefore tested effects of conditioning lesions plus local delivery of NT3 and BDNF on AST axon regeneration stimulated in the presence or absence of astrocyte scar formation (Fig. 5 and Extended Data Figs 7–9). Because cell grafts modify astrocytic scars[35] and provide permissive substrates for regrowing axons, we delivered NT3 and BDNF via synthetic hydrogel depots that do not modify astrocytic scar formation and provide prolonged neurotrophin delivery[37–39] (Fig. 5d; Supplementary Information).

No AST fibres regrew past astrocytic scars into lesion cores after SCI alone or with hydrogel without growth factors (Fig. 5a, j and

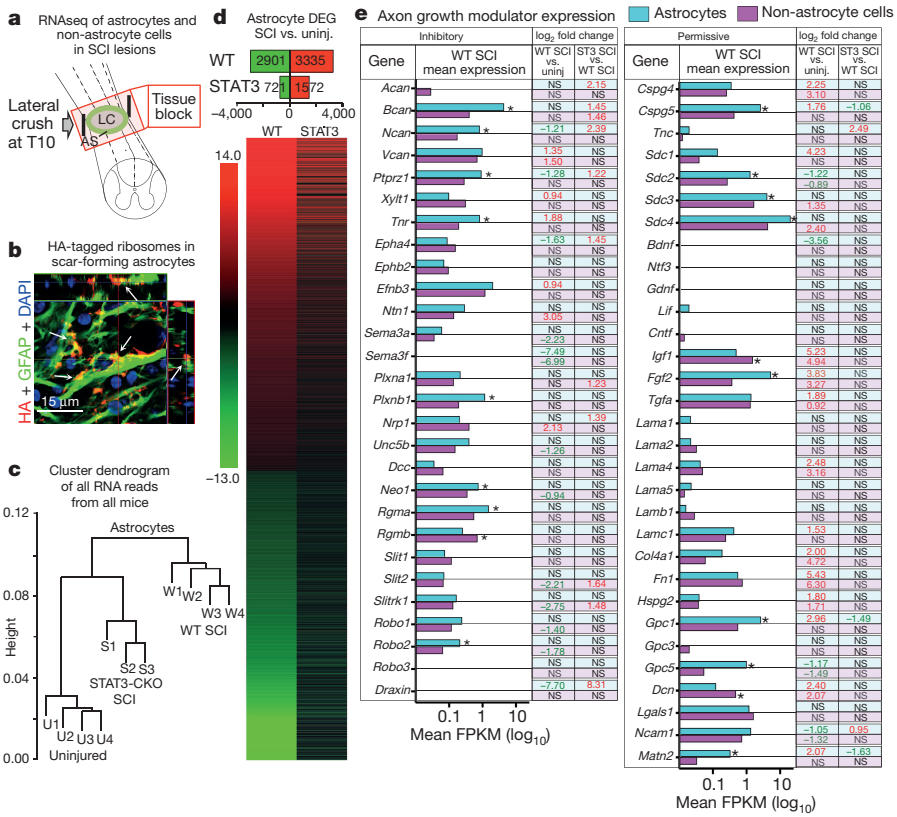


Figure 4 | Genomic dissection of astrocytes and non-astrocyte cells in SCI lesions. a, Schematic of tissue harvested for RNA sequencing (RNAseq). AS, astrocytic scar; LC, lesion core. **b**, Transgenically-targeted haemagglutinin (HA)-tagged ribosome clusters (arrows) among GFAP filaments in scar-forming astrocytes (see Extended Data Fig. 3a). **c**, Unsupervised hierarchical clustering dendrogram based on Pearson correlations shows that transcriptome profiles of STAT3-CKO SCI astrocytes cluster nearer to (and are more similar to) those of uninjured as compared to wild-type (WT) SCI astrocytes. **d**, Numbers and heat map of significantly differentially expressed genes (DEGs) in wild-type astrocytes after SCI and the comparative differential expression profile of that specific cohort of genes in STAT3-CKO SCI astrocytes. Differential expression relative to uninjured, false discovery rate (FDR) < 0.1. **e**, Histogram of astrocyte and non-astrocyte expression of axon-growth-inhibitory or -permissive molecules after SCI shown as mean FPKM (fragments per kilobase of transcript sequence per million mapped fragments). Asterisks indicate significant differences between astrocytes and non-astrocytes, FDR < 0.1. Numbers show significant $\log_2$ fold differences. Red, upregulated; green, downregulated; NS, non-significant; ST3 SCI, STAT3-CKO SCI.

Extended Data Fig. 7c). Small numbers of AST axons regrew into lesion cores in mice with SCI plus conditioning lesions alone, or SCI plus NT3 and BDNF without conditioning lesions (Fig. 5j, k and Extended Data Fig. 7c). By contrast, mice receiving both conditioning lesions plus NT3 and BDNF exhibited robust axon regrowth through and beyond astrocytic scars, and the number of axon intercepts counted at lesion centres averaged over 45% of that in intact sensory tracts 3 mm proximal to lesions (Fig. 5b, j, k and Extended Data Fig. 7c). Remarkably, these stimulated AST axons regrew profusely through, and past, dense astrocyte scars, and in spite of substantial CSPG (Fig. 5b, e–g, Extended Data Figs 7b, c and 8). AST axons stimulated to regrow in SCI lesions were thin and uniformly tracked along laminin surfaces, turning and even reversing direction along these surfaces as expected of regenerating axons[40], whereas AST axons in intact gracile–cuneate tracts were coarsely beaded and were not in direct contact with laminin (Fig. 5h and Extended Data Fig. 9a–f). Hydrogel delivery of anti-CD29 laminin-integrin function-blocking antibodies[36] together with NT3 and BDNF significantly reduced AST axon regrowth in lesion cores by 73%, demonstrating that laminin interactions were critical for stimulated AST axon regrowth (Fig. 5i, j, Extended Data Fig. 9g–i and Supplementary Information). Lastly, preventing astrocyte scar formation did not augment AST axon regrowth stimulated by conditioning lesions plus NT3 and BDNF, but instead completely prevented, or significantly attenuated, axon regrowth in TK + GCV mice and STAT3-CKO mice, respectively (Fig. 5c, j, k), demonstrating that astrocytic scar formation aids, rather than inhibits appropriately stimulated AST axon regeneration after SCI.

Discussion

Our findings show that contrary to prevailing dogma, astrocytic scar formation is not a principal cause for the failure of injured mature CNS axons to regrow across severe CNS lesions and that scar-forming astrocytes permit and support robust amounts of appropriately stimulated CNS axon regeneration. Although our observations with stimulated AST axon regeneration needs extension to other axonal systems, our findings are consistent with evidence that: (i) astrocytes can support growth of different CNS axons *in vivo* during development[41,42] or after mature CNS injury[19,43]; (ii) genetic activation of axonal growth programs by mature neurons leads to axon regeneration across CNS lesions only when scar-forming astrocyte bridges are present[2,44]; and (iii) grafts of progenitor-derived astrocytes support axon regeneration through non-neural SCI lesion cores[45,46].

The predominant mechanistic proposal for astrocytic scar inhibition of axon regeneration is CSPG production[9]. However, specific CSPGs can support or repel axon growth (see Extended Data Table 1). We show that scar-forming astrocytes upregulate growth supportive CSPG4 and CSPG5, and that both astrocytes and non-astrocytes in SCI lesions express multiple axon-growth-supportive molecules, including laminins. Axon growth and guidance depend both on intrinsic growth potential[2,3,31,32,47] and on a balance of extrinsic regulatory cues that modulate one another's effects on growth cones[25]. Our findings show that sustained delivery of required axon-specific growth factors not adequately expressed in SCI lesions, combined with activation of neuron-intrinsic growth programs, can stimulate robust regrowth of transected axons along specifically required supportive matrix cues in

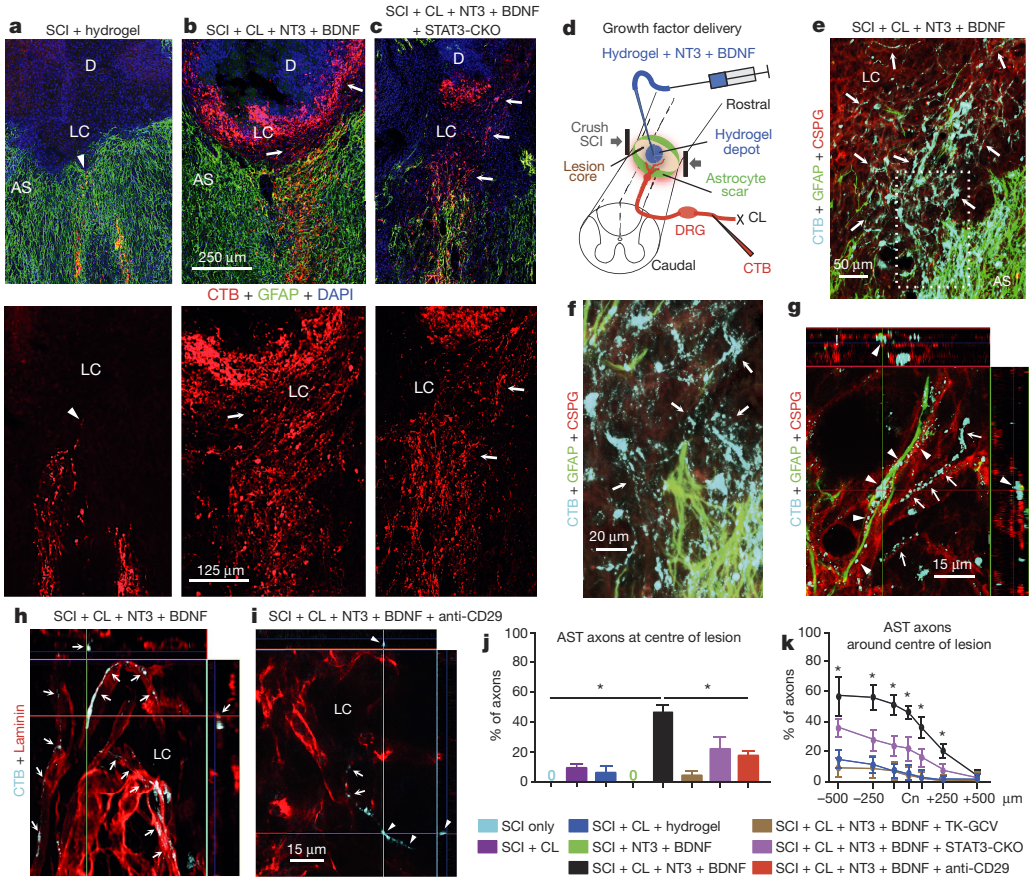


Figure 5 | Robust regrowth of AST axons can be stimulated after wild-type SCI and is significantly attenuated by preventing astrocytic scar formation. a–c, Top, AST axons (choleratoxin B tracing) plus GFAP immunohistochemistry. Bottom, AST axons alone. AS, astrocytic scar; D, hydrogel depot; LC, lesion core. **a**, Wild-type mouse, SCI and hydrogel only (no growth factors). Arrowhead denotes most rostrally penetrating axons that do not pass beyond AS. **b**, Wild-type mouse, SCI plus conditioning lesion (CL) and hydrogel depot (D) with NT3 + BDNF. Arrows denote robust regrowth of AST axons past the scar into the lesion core and along, but not into, the depot that releases NT3 + BDNF but that provides no adhesive matrix. **c**, STAT3-CKO mouse, SCI plus conditioning lesions and NT3 + BDNF depot. Arrows denote regrowth of AST axons into the lesion core. **d**, Experiment summary schematic. CTB, choleratoxin B; DRG, dorsal root ganglion. **e–g**, Wild-type mice, AST plus GFAP and CSPG (CS56) immunohistochemistry. Box in **e** is shown in **f**. **e**, **f**, Arrows denote robust regrowth of stimulated AST axons past the astrocytic scar into the lesion core through CSPG. **g**, Regrowing AST axons track along CSPG-positive GFAP-negative structures (arrows) or along CSPG-positive GFAP-positive astrocyte processes (arrowheads) (see Extended Data Figs 7 and 8 for single channel images of **e–g**). **h**, **i**, AST axons plus laminin immunohistochemistry. **h**, **i**, Arrows indicate regrowing stimulated AST axons tracking along laminin. **i**, Arrowheads indicate stimulated AST axons exposed to anti-CD29 antibody and failing to maintain contact with laminin. **j**, **k**, Numbers of AST axons at SCI lesion centre (Cn) (**j**) or on either side (**k**) expressed as a percentage of all axons 3 mm proximal. $n = 5$; $^{*}P < 0.05$ versus all other groups, $^{\#}P < 0.05$ versus all groups except STAT3-CKO (ANOVA with Newman–Keuls).

spite of the presence of inhibitory cues, and that this stimulated axon regrowth can occur either in direct contact with scar-forming astrocyte processes or independently of astrocyte processes. These observations provide direct evidence that the requirements for achieving axon regeneration across severe CNS lesions, where transected axons lack intrinsic and extrinsic conditions for long-distance regrowth, are fundamentally different from requirements to achieve local neurite outgrowth in intact but reactive perilesional grey matter, where conditions compatible with axon terminal growth and remodelling are present and where blocking inhibitory regulators such as CSPGs and others produced by astrocytes and other cells may be sufficient to promote axon sprouting that might also improve function[9,15,24,28,40,48]. Our findings have important implications for CNS repair strategies by demonstrating that rather than being hostile to axon growth, newly generated immature scar-forming astrocytes derived after SCI from endogenous progenitors[18,41,44], and potentially from grafted progenitors[45,46], aid axon regeneration and may represent exploitable bridges for regrowing axons across severe CNS lesions.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 23 October 2015; accepted 26 February 2016.
Published online 30 March; corrected online 13 April 2016 (see full-text HTML version for details).

1. Ramón y Cajal, S. *Degeneration and Regeneration of the Nervous System* (Oxford Univ. Press, 1928).
2. Sun, F. *et al.* Sustained axon regeneration induced by co-deletion of PTEN and SOCS3. *Nature* **480,** 372–375 (2011).
3. Liu, K., Tedeschi, A., Park, K. K. & He, Z. Neuronal intrinsic mechanisms of axon regeneration. *Annu. Rev. Neurosci.* **34,** 131–152 (2011).
4. Richardson, P. M., McGuinness, U. M. & Aguayo, A. J. Axons from CNS neurons regenerate into PNS grafts. *Nature* **284,** 264–265 (1980).
5. David, S. & Aguayo, A. J. Axonal elongation into peripheral nervous system "bridges" after central nervous system injury in adult rats. *Science* **214,** 931–933 (1981).
6. Schwab, M. E. Functions of Nogo proteins and their receptors in the nervous system. *Nature Rev. Neurosci.* **11,** 799–811 (2010).
7. Harel, N. Y. & Strittmatter, S. M. Can regenerating axons recapitulate developmental guidance during recovery from spinal cord injury? *Nature Rev. Neurosci.* **7,** 603–616 (2006).
8. Klapka, N. & Muller, H. W. Collagen matrix in spinal cord injury. *J. Neurotrauma* **23,** 422–435 (2006).
9. Silver, J. & Miller, J. H. Regeneration beyond the glial scar. *Nature Rev. Neurosci.* **5,** 146–156 (2004).
10. Windle, W. F., Clemente, C. D. & Chambers, W. W. Inhibition of formation of a glial barrier as a means of permitting a peripheral nerve to grow into the brain. *J. Comp. Neurol.* **96,** 359–369 (1952).

11. Windle, W. F. Regeneration of axons in the vertebrate central nervous system. *Physiol. Rev.* **36,** 427–440 (1956).
12. Liuzzi, F. J. & Lasek, R. J. Astrocytes block axonal regeneration in mammals by activating the physiological stop pathway. *Science* **237,** 642–645 (1987).
13. Burda, J. E. & Sofroniew, M. V. Reactive gliosis and the multicellular response to CNS damage and disease. *Neuron* **81,** 229–248 (2014).
14. Sofroniew, M. V. Astrocyte barriers to neurotoxic inflammation. *Nature Rev. Neurosci.* **16,** 249–263 (2015).
15. Bush, T. G. *et al.* Leukocyte infiltration, neuronal degeneration and neurite outgrowth after ablation of scar-forming, reactive astrocytes in adult transgenic mice. *Neuron* **23,** 297–308 (1999).
16. Faulkner, J. R. *et al.* Reactive astrocytes protect tissue and preserve function after spinal cord injury. *J. Neurosci.* **24,** 2143–2155 (2004).
17. Herrmann, J. E. *et al.* STAT3 is a critical regulator of astrogliosis and scar formation after spinal cord injury. *J. Neurosci.* **28,** 7231–7243 (2008).
18. Wanner, I. B. *et al.* Glial scar borders are formed by newly proliferated, elongated astrocytes that interact to corral inflammatory and fibrotic cells via STAT3-dependent mechanisms after spinal cord injury. *J. Neurosci.* **33,** 12870–12886 (2013).
19. Lee, J. K. *et al.* Combined genetic attenuation of myelin and semaphorin-mediated growth inhibition is insufficient to promote serotonergic axon regeneration. *J. Neurosci.* **30,** 10899–10904 (2010).
20. Hawthorne, A. L. *et al.* The unusual response of serotonergic neurons after CNS injury: lack of axonal dieback and enhanced sprouting within the inhibitory environment of the glial scar. *J. Neurosci.* **31,** 5605–5616 (2011).
21. Buch, T. *et al.* A Cre-inducible diphtheria toxin receptor mediates cell lineage ablation after toxin administration. *Nature Methods* **2,** 419–426 (2005).
22. Avnur, Z. & Geiger, B. Immunocytochemical localization of native chondroitin-sulfate in tissues and cultured cells using specific monoclonal antibody. *Cell* **38,** 811–822 (1984).
23. Mikami, T. & Kitagawa, H. Biosynthesis and function of chondroitin sulfate. *Biochim. Biophys. Acta* **1830,** 4719–4733 (2013).
24. Mironova, Y. A. & Giger, R. J. Where no synapses go: gatekeepers of circuit remodeling and synaptic strength. *Trends Neurosci.* **36,** 363–373 (2013).
25. Lin, A. C. & Holt, C. E. Local translation and directional steering in axons. *EMBO J.* **26,** 3729–3736 (2007).
26. Sanz, E. *et al.* Cell-type-specific isolation of ribosome-associated mRNA from complex tissues. *Proc. Natl Acad. Sci. USA* **106,** 13939–13944 (2009).
27. Zamanian, J. L. *et al.* Genomic analysis of reactive astrogliosis. *J. Neurosci.* **32,** 6391–6410 (2012).
28. Lang, B. T. *et al.* Modulation of the proteoglycan receptor PTPsigma promotes recovery after spinal cord injury. *Nature* **518,** 404–408 (2015).
29. Yamaguchi, Y. Lecticans: organizers of the brain extracellular matrix. *Cell. Mol. Life Sci.* **57,** 276–289 (2000).
30. Miller, G. M. & Hsieh-Wilson, L. C. Sugar-dependentmodulation of neuronal development, regeneration, and plasticity by chondroitin sulfate proteoglycans. *Exp. Neurol.* **274,** 115–125 (2015).
31. Goldberg, J. L. *et al.* Retinal ganglion cells do not extend axons by default: promotion by neurotrophic signaling and electrical activity. *Neuron* **33,** 689–702 (2002).
32. Richardson, P. M. & Issa, V. M. Peripheral injury enhances central regeneration of primary sensory neurones. *Nature* **309,** 791–793 (1984).
33. Neumann, S. & Woolf, C. J. Regeneration of dorsal column fibers into and beyond the lesion site following adult spinal cord injury. *Neuron* **23,** 83–91 (1999).
34. Omura, T. *et al.* Robust Axonal regeneration occurs in the injured CAST/Ei mouse CNS. *Neuron* **86,** 1215–1227 (2015).
35. Alto, L. T. *et al.* Chemotropic guidance facilitates axonal regeneration and synapse formation after spinal cord injury. *Nature Neurosci.* **12,** 1106–1113 (2009).
36. Plantman, S. *et al.* Integrin-laminin interactions controlling neurite outgrowth from adult DRG neurons *in vitro*. *Mol. Cell. Neurosci.* **39,** 50–62 (2008).
37. Nowak, A. P. *et al.* Rapidly recovering hydrogel scaffolds from self-assembling diblock copolypeptide amphiphiles. *Nature* **417,** 424–428 (2002).
38. Yang, C. Y. *et al.* Biocompatibility of amphiphilic diblock copolypeptide hydrogels in the central nervous system. *Biomaterials* **30,** 2881–2898 (2009).
39. Song, B. *et al.* Sustained local delivery of bioactive nerve growth factor in the central nervous system via tunable diblock copolypeptide hydrogel depots. *Biomaterials* **33,** 9105–9116 (2012).
40. Tuszynski, M. H. & Steward, O. Concepts and methods for the study of axonal regeneration in the CNS. *Neuron* **74,** 777–791 (2012).
41. Brosius Lutz, A. & Barres, B. A. Contrasting the glial response to axon injury in the central and peripheral nervous systems. *Dev. Cell* **28,** 7–17 (2014).
42. Mason, C. A., Edmondson, J. C. & Hatten, M. E. The extending astroglial process: development of glial cell shape, the growing tip, and interactions with neurons. *J. Neurosci.* **8,** 3124–3134 (1988).
43. Kawaja, M. D. & Gage, F. H. Reactive astrocytes are substrates for the growth of adult CNS axons in the presence of elevated levels of nerve growth factor. *Neuron* **7,** 1019–1030 (1991).
44. Zukor, K. *et al.* Short hairpin RNA against PTEN enhances regenerative growth of corticospinal tract axons after spinal cord injury. *J. Neurosci.* **33,** 15350–15361 (2013).
45. Shih, C. H., Lacagnina, M., Leuer-Bisciotti, K. & Proschel, C. Astroglial-derived periostin promotes axonal regeneration after spinal cord injury. *J. Neurosci.* **34,** 2438–2443 (2014).
46. Zhang, S. *et al.* Thermoresponsive copolypeptide hydrogel vehicles for CNS cell delivery. *ACS Biomater. Sci. Eng.* **1,** 705–717 (2015).
47. Ruschel, J. *et al.* Systemic administration of epothilone B promotes axon regeneration after spinal cord injury. *Science* **348,** 347–352 (2015).
48. Cafferty, W. B., McGee, A. W. & Strittmatter, S. M. Axonal growth therapeutics: regeneration or sprouting or plasticity? *Trends Neurosci.* **31,** 215–220 (2008).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank D. W. Bergles for the NG2 antibody, and the Microscopy Core Resource of the UCLA Broad Stem Cell Research Center-CIRM Laboratory. This work was supported by the US National Institutes of Health (NS057624 and NS084030 to M.V.S.; P30 NS062691 to G.C. and NS060677, MH099559A, MH104069 to B.S.K.), and the Dr. Miriam and Sheldon G. Adelson Medical Foundation (M.V.S. and T.J.D.), and Wings for Life (M.V.S.).

Author Contributions M.A.A., J.E.B., B.S.K., T.J.D. and M.V.S. designed experiments; M.A.A., J.E.B., Y.R. and Y.A. conducted experiments; M.A.A., J.E.B., Y.A., T.M.O., R.K., G.C. and M.V.S. analysed data. M.A.A., J.E.B., T.M.O., B.S.K, T.J.D. and M.V.S. prepared the manuscript.

Author Information Raw and normalized genomic data have been deposited in the NCBI Gene Expression Omnibus and are accessible through accession number GSE76097 and via a searchable, open-access website https://astrocyte.rnaseq.sofroniewlab.neurobio.ucla.edu. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.V.S. (sofroniew@mednet.ucla.edu).

METHODS

Mice. All non-transgenic, transgenic and control mice used in this study were derived from in house breeding colonies backcrossed > 12 generations onto C57/BL6 backgrounds. All mice used were young adult females between two and four months old at the time of spinal cord injury. All transgenic mice used have been previously well characterized or are the progeny of crossing well-characterized lines: (1) mGFAP-TK transgenic mice line 7.1[15,16,49]; (2) mGFAP-Cre-STAT3-*loxP* mice generated by crossing STAT3-*loxP* mice with *loxP* sites flanking exon 22 of the STAT3 gene[50] with mGFAP-Cre mice line 73.12[17,18]; (3) *loxP*-STOP-*loxP*-DTR (diphtheria toxin receptor) mice[21]; (4) mGFAP-Cre-RiboTag mice generated by crossing mice with *loxP*-STOP-*loxP*-*Rpl22*-HA (RiboTag)[26] with mGFAP-Cre mice line 73.12[17,18]; (5) *loxP*-STOP-*loxP*-tdTomato reporter mice[51]. All mice were housed in a 12-h light/dark cycle in a specific-pathogen-free facility with controlled temperature and humidity and were allowed free access to food and water. All experiments were conducted according to protocols approved by the Animal Research Committee of the Office for Protection of Research Subjects at University of California, Los Angeles.

Surgical procedures. All surgeries were performed under general anaesthesia with isoflurane in oxygen-enriched air using an operating microscope (Zeiss, Oberkochen, Germany), and rodent stereotaxic apparatus (David Kopf, Tujunga, CA). Laminectomy of a single vertebra was performed and severe crush spinal cord injuries (SCI) were made at the level of T10 using No. 5 Dumont forceps (Fine Science Tools, Foster City, CA) without spacers and with a tip width of 0.5 mm to completely compress the entire spinal cord laterally from both sides for 5 s[16–18]. For pre-conditioning lesions, sciatic nerves were transected and ligated one week before SCI. Hydrogels were injected stereotaxically into the centre of SCI lesions 0.6 mm below the surface at 0.2 μl per minute using glass micropipettes (ground to 50–100 μm tips) connected via high-pressure tubing (Kopf) to 10-μl syringes under control of microinfusion pumps, two days after SCI[52]. Tract tracing was performed by injection of biotinylated dextran amine 10,000 (BDA, Invitrogen) 10% wt/vol in sterile saline injected 4 × 0.4 μl into the left motor cerebral cortex 14 days before perfusion to visualize corticospinal tract (CST) axons, or choleratoxin B (CTB) (List Biological Laboratory, Campbell, CA) 1 μl of 1% wt/vol in sterile water injected into both sciatic nerves three days before perfusion to visualize ascending sensory tract (AST) axons[33]. AAV2/5-GfaABC1D-Cre (see below) was injected either 3 or 6 × 0.4 μl (1.29×10^{13} gc ml^{-1} in sterile saline) into and on either side of mature SCI lesions two weeks after SCI, or into uninjured spinal cord after T10 laminectomy. All animals received analgesic before wound closure and every 12 h for at least 48 h post-injury. Animals were randomly assigned numbers and evaluated thereafter blind to genotype and experimental condition.

AAV2/5-GfaABC1D-Cre. Adeno-associated virus 2/5 (AAV) vector with a minimal GFAP promoter (AAV2/5 GfaABC1D) was used to target Cre-recombinase expression selectively to astrocytes[53–55].

Hydrogel with growth factors and antibodies. Diblock co-polypeptide hydrogel (DCH) $K_{180}L_{20}$ was fabricated, tagged with blue fluorescent dye (AMCA-X) and loaded with growth factor and antibody cargoes as described[38,39,52]. Cargo molecules comprised: human recombinant NT3 and BDNF were gifts (Amgen, Thousand Oaks, CA, (NT3 Lot#2200F4; BDNF Lot#2142F5A) or were purchased from PeproTech (Rocky Hill, NJ; NT3 405-03, Lot#060762; BDNF 405-02 Lot#071161). Function blocking anti-CD29 mouse monoclonal antibody was purchased from BD Bioscience (San Diego, CA) as a custom order at 10.25 mg ml^{-1} (product #BP555003; lot#S03146). Freeze dried $K_{180}L_{20}$ powder was reconstituted on to 3.0% or 3.5% wt/vol basis in sterile PBS without cargo or with combinations of NT3 (1.0 μg μl^{-1}), BDNF (0.85 μg μl^{-1}) and anti-CD29 (5 μg μl^{-1}). DCH mixtures were prepared to have G′ (storage modulus at 1 Hz) between 75 and 100 Pascal (Pa), somewhat below that of mouse brain at 200 Pa (refs 38,39).

Ganciclovir (GCV), BrdU or DT injections. GCV (Cytovene-IV Hoffman LaRoche, Nutley, NJ), 25 mg kg^{-1} per day dissolved in sterile physiological saline was administered as single daily subcutaneous injections starting immediately after surgery and continued for the first 7 days after SCI. Bromodeoxyuridine (BrdU, Sigma), 100 mg kg^{-1} per day dissolved in saline plus 0.007 M NaOH, was administered as single daily intraperitoneal injections on days 2 through 7 after SCI. Diphtheria toxin A (DT, Sigma #DO564) 100 ng in 100 μl sterile saline was administered twice daily as intraperitoneal injections for ten days starting three weeks after injection of AAV2/5-GfaABC1D-Cre to *loxP*-DTR mice (which was 5 weeks after SCI) (see timeline in Extended Data Fig. 1d).

Hindlimb locomotor evaluation, animal inclusion criteria, randomization and blinding. Two days after SCI, all mice were evaluated in open field and mice exhibiting any hindlimb movements were not studied further. Mice that passed this pre-determined inclusion criterion were randomized into experimental groups for further treatments and were thereafter evaluated blind to their experimental condition. At 3, 7, 14 days and then weekly after SCI, hindlimb movements were scored using a simple six-point scale in which 0 is no movement and 5 is normal walking[17].

Histology and immunohistochemistry. After terminal anaesthesia by barbiturate overdose mice were perfused transcardially with 10% formalin (Sigma). Spinal cords were removed, post-fixed overnight, and cryoprotected in buffered 30% sucrose for 48 h. Frozen sections (30 μm horizontal) were prepared using a cryostat microtome (Leica) and processed for immunofluorescence as described[16–18]. Primary antibodies were: rabbit anti-GFAP (1:1,000; Dako, Carpinteria, CA); rat anti-GFAP (1:1,000, Zymed Laboratories); goat anti-CTB (1:1,000, List Biology Lab); rabbit anti-5HT (1:2,000, Immunostar); goat anti-5HT (1:1,000, Immunostar); mouse anti-CSPG[22] (1:100, Sigma); rabbit-anti haemagglutinin (HA) (1:500 Sigma); mouse-anti HA (1:3,000 Covance); sheep anti-BrdU (1:6,000, Maine Biotechnology Services, Portland, ME); rabbit anti-laminin (1:80, Sigma, Saint Louis, MO); guinea pig anti-NG2 (CSPG4) (E. G. Hughes and D. W. Bergles[56], Baltimore, MA); goat anti-aggrecan (1:200, NOVUS); rabbit anti-brevican (1:300, NOVUS); mouse anti-neurocan (1:300, Milipore); mouse anti-phosphacan (1:500, Sigma); goat anti-versican (1:200, NOVUS); rabbit anti-neurglycan C (CSPG5) (1:200, NOVUS). Fluorescence secondary antibodies were conjugated to: Alexa 488 (green) or Alexa 350 (blue) (Molecular Probes), or to Cy3 (550, red) or Cy5 (649, far red) all from (Jackson Immunoresearch Laboratories). Mouse primary antibodies were visualized using the Mouse-on-Mouse detection kit (M.O.M., Vector). BDA tract-tracing was visualized with streptavidin-HRP plus TSB Fluorescein green or Tyr-Cy3 (Jackson Immunoresearch Laboratories). Nuclear stain: 4′,6′-diamidino-2-phenylindole dihydrochloride (DAPI; 2 ng ml^{-1}; Molecular Probes). Sections were coverslipped using ProLong Gold anti-fade reagent (InVitrogen, Grand Island, NY). Sections were examined and photographed using deconvolution fluorescence microscopy and scanning confocal laser microscopy (Zeiss, Oberkochen, Germany).

Axon quantification. Axons labelled by tract tracing or immunohistochemistry were quantified using image analysis software (NeuroLucida, MicroBrightField, Williston, VT) operating a computer-driven microscope regulated in the *x*, *y* and *z* axes (Zeiss) by observers blind to experimental conditions. Using NeuroLucida, lines were drawn across horizontal spinal cord sections at SCI lesion centres and at regular distances on either side (Fig. 1a) and the number of axons intercepting lines was counted at 63× magnification under oil immersion by observers blind to experimental conditions. Similar lines were drawn and axons counted in intact axon tracts 3 mm proximal to SCI lesions and the numbers of axon intercepts in or near lesions were expressed as percentages of axons in the intact tracts in order to control for potential variations in tract-tracing efficacy or intensity of immunohistochemistry among animals. Two sections at the level of the CST or AST, and three sections through the middle of the cord for 5HT, were counted per mouse and expressed as total intercepts per location per mouse. To determine efficacy of axon transection after SCI, we examined labelling 3 mm distal to SCI lesion centres, with the intention of eliminating mice that had labelled axons at this location on grounds that these mice may have had incomplete lesions. However, all mice that had met the strict behavioural inclusion criterion of no hindlimb movements two days after severe crush SCI, exhibited no detectable axons 3 mm distal to SCI lesions regardless of treatment group.

Quantification of immunohistochemically stained areas. Sections stained for GFAP, CSPG or laminin were photographed using constant exposure settings. Single-channel immunofluorescence images were converted to black and white and thresholded (Fig. 1d and Extended Data Fig. 2b) and the amount of stained area measured in different tissue compartments using NIH ImageJ software. Areas are shown in graphs as mean values plus or minus standard error of the means (s.e.m.).

Statistics, power calculations and group sizes. Statistical evaluations of repeated measures were conducted by ANOVA with post hoc, independent pairwise analysis as per Newman-Keuls (Prism, GraphPad, San Diego, CA). Power calculations were performed using G*Power Software v3.1.9.2 (ref. 57). For quantification of histologically derived neuroanatomical outcomes such as numbers of axons or percentage of area stained for GFAP or CSPG, group sizes were used that were calculated to provide at least 80% power when using the following parameters: probability of type I error (α) = 0.05, a conservative effect size of 0.25, 2–8 treatment groups with multiple measurements obtained per replicate. Using Fig. 5j as an example, evaluation of $n = 5$ biological replicates (with multiple measurements per replicate) in each of 8 treatment groups provided greater than 88% power.

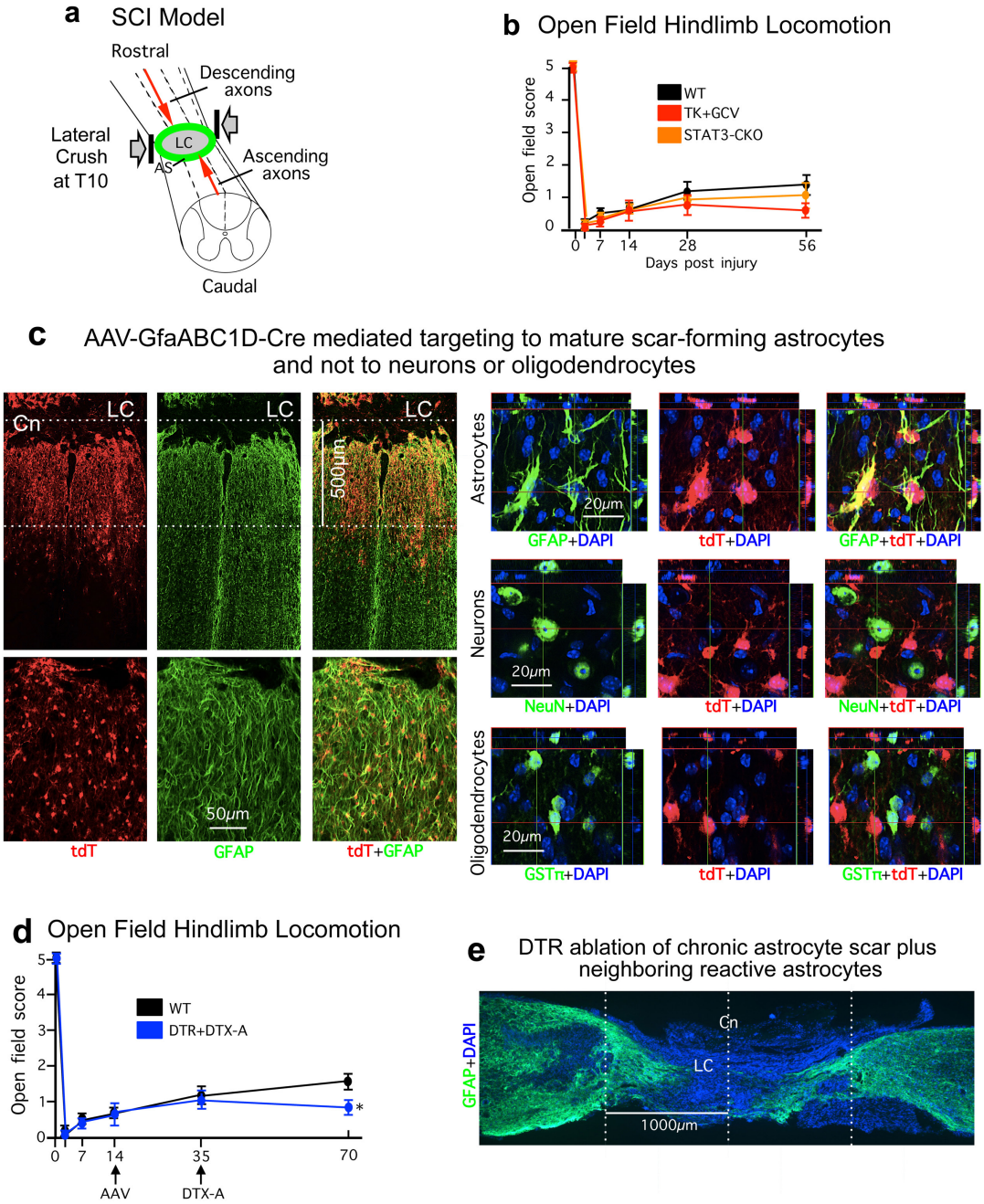
Dot blot. For dot blot immunoassay of chondroitin sulfate proteoglycans (CSPG), spinal cord tissue blocks were lysed and homogenized in standard RIPA (radio-immunoprecipitation assay) buffer. LDS (lithium dodecyl sulfate) buffer (Life Technologies) was added to the post-mitochondrial supernatant and 2 μl containing 2 μg μl^{-1} protein was spotted onto a nitrocellulose membrane

(Life Technologies), set to dry and incubated overnight with mouse anti-chondroitin sulfate antibody (CS56, 1:1000, Sigma Aldrich), an IgM-monoclonal antibody that detects glyco-moieties of all CSPGs[22]. CS56 immunoreactivity was detected on X-ray film with alkaline phosphatase-conjugated secondary antibody and chemiluminescent substrate (Life Technologies). Densitometry measurements of CS56 immunoreactivity were obtained using ImageJ software (NIH) and normalized to total protein (Poncau S) density[58]. Densities are shown in graphs as mean values plus or minus standard error of the means (s.e.m.).

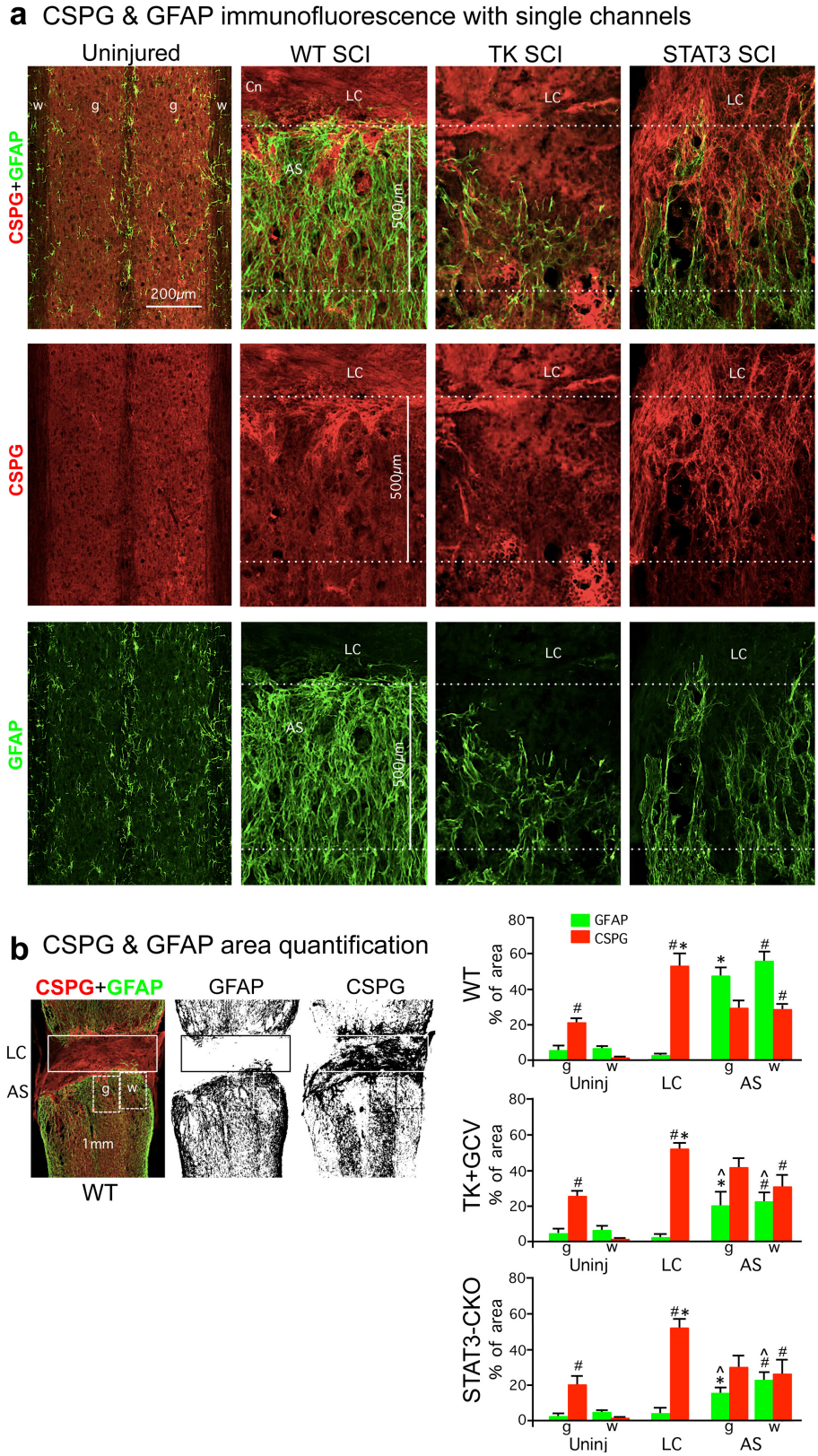
Isolation, sequencing and analysis of RNA from astrocytes and non-astrocyte cells. Two weeks after SCI, spinal cords of wild-type control (GFAP-RiboTag) and STAT3-CKO (GFAP-STAT3CKO-RiboTag) mice were rapidly dissected out of the spinal canal. The central 3 mm of the lower thoracic lesion including the lesion core and 1 mm rostral and caudal were then rapidly removed and snap frozen in liquid nitrogen. Haemagglutinin (HA) immunoprecipitation (HA-IP) of astrocyte ribosomes and ribosome-associated mRNA (ramRNA) was carried out as described[26]. The non-precipitated flow-through (FT) from each IP sample was collected for analysis of non-astrocyte total RNA. HA and FT samples underwent on-column DNA digestion using the RNase-Free Dnase Set (Qiagen) and RNA purified with the RNeasy Micro kit (Qiagen). Integrity of the eluted RNA was analysed by a 2100 Bioanalyzer (Agilent) using the RNA Pico chip, mean sample RIN = 8.0 ± 0.95. RNA concentration determined by RiboGreen RNA Assay kit (Life Technologies). cDNA was generated from 5 ng of IP or FT RNA using the Nugen Ovation 2 RNA-Seq Sytstem V2 kit (Nugen). 1 μg of cDNA was fragmented using the Covaris M220. Paired-end libraries for multiplex sequencing were generated from 300 ng of fragmented cDNA using the Apollo 324 automated library preparation system (Wafergen Biosystems) and purified with Agencourt AMPure XP beads (Beckman Coulter). All samples were analysed by an Illumina NextSeq 500 Sequencer (Illumina) using 75-bp paired-end sequencing. Reads were quality controlled using in-house scripts including picard-tools, mapped to the reference mm10 genome using STAR[59], and counted using HT-seq[60] with mm10 refSeq as reference, and genes were called differentially expressed using edgeR[61]. Individual gene expression levels in the Fig. 4e histogram are shown as mean FPKM (fragments per kilobase of transcript sequence per million mapped fragments). Additional details of differential expression analysis are described in the legends of Fig. 4 and Extended Data Figs 3 and 4. Raw and normalized data have been deposited in the NCBI Gene Expression Omnibus and are accessible through accession number GSE76097. To ensure the widespread distribution of these datasets, we have created a user-friendly website that enables searching for individual genes of interest https://astrocyte.rnaseq.sofroniewlab.neurobio.ucla.edu.

49. Bush, T. G. *et al.* Fulminant jejuno-ileitis following ablation of enteric glia in adult transgenic mice. *Cell* **93,** 189–201 (1998).
50. Takeda, K. *et al.* Stat3 activation is responsible for IL-6-dependent T cell proliferation through preventing apoptosis: generation and characterization of T cell-specific Stat3-deficient mice. *J. Immunol.* **161,** 4652–4660 (1998).
51. Madisen, L. *et al.* A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. *Nature Neurosci.* **13,** 133–140 (2010).
52. Zhang, S. *et al.* Tunable diblock copolypeptide hydrogel depots for local delivery of hydrophobic molecules in healthy and injured central nervous system. *Biomaterials* **35,** 1989–2000 (2014).
53. Shigetomi, E. *et al.* Imaging calcium microdomains within entire astrocyte territories and endfeet with GCaMPs expressed using adeno-associated viruses. *J. Gen. Physiol.* **141,** 633–647 (2013).
54. Jiang, R., Haustein, M. D., Sofroniew, M. V. & Khakh, B. S. Imaging intracellular Ca^{2+} signals in striatal astrocytes from adult mice using genetically-encoded calcium indicators. *J. Vis. Exp.* **93,** e51972 (2014).
55. Tong, X. *et al.* Astrocyte Kir4.1 ion channel deficits contribute to neuronal dysfunction in Huntington's disease model mice. *Nature Neurosci.* **17,** 694–703 (2014).
56. Kang, S. H. *et al.* Degeneration and impaired regeneration of gray matter oligodendrocytes in amyotrophic lateral sclerosis. *Nature Neurosci.* **16,** 571–579 (2013).
57. Faul, F., Erdfelder, E., Lang, A. G. & Buchner, A. G. *Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* **39,** 175–191 (2007).
58. Romero-Calvo, I. *et al.* Reversible Ponceau staining as a loading control alternative to actin in western blots. *Anal. Biochem.* **401,** 318–320 (2010).
59. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
60. Anders, S., Pyl, P. T. & Huber, W. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31,** 166–169 (2015).
61. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).
62. Friedlander, D. R. *et al.* The neuronal chondroitin sulfate proteoglycan neurocan binds to the neural cell adhesion molecules Ng-CAM/L1/NILE and N-CAM, and inhibits neuronal adhesion and neurite outgrowth. *J. Cell Biol.* **125,** 669–680 (1994).
63. Sango, K. *et al.* Phosphacan and neurocan are repulsive substrata for adhesion and neurite extension of adult rat dorsal root ganglion neurons in vitro. *Exp. Neurol.* **182,** 1–11 (2003).
64. Hurtado, A., Podinin, H., Oudega, M. & Grimpe, B. Deoxyribozyme-mediated knockdown of xylosyltransferase-1 mRNA promotes axon growth in the adult rat spinal cord. *Brain* **131,** 2596–2605 (2008).
65. Becker, C. G., Schweitzer, J., Feldner, J., Becker, T. & Schachner, M. Tenascin-R as a repellent guidance molecule for developing optic axons in zebrafish. *J. Neurosci.* **23,** 6232–6237 (2003).
66. Dickson, B. J. Molecular mechanisms of axon guidance. *Science* **298,** 1959–1964 (2002).
67. Masuda, T. *et al.* Netrin-1 acts as a repulsive guidance cue for sensory axonal projections toward the spinal cord. *J. Neurosci.* **28,** 10380–10385 (2008).
68. Winberg, M. L. *et al.* Plexin A is a neuronal semaphorin receptor that controls axon guidance. *Cell* **95,** 903–916 (1998).
69. Hu, H., Marton, T. F. & Goodman, C. S. Plexin B mediates axon guidance in Drosophila by simultaneously inhibiting active Rac and enhancing RhoA signaling. *Neuron* **32,** 39–51 (2001).
70. He, Z. & Tessier-Lavigne, M. Neuropilin is a receptor for the axonal chemorepellent Semaphorin III. *Cell* **90,** 739–751 (1997).
71. Lu, X. *et al.* The netrin receptor UNC5B mediates guidance events controlling morphogenesis of the vascular system. *Nature* **432,** 179–186 (2004).
72. Keino-Masu, K. *et al. Deleted in Colorectal Cancer* (DCC) encodes a netrin receptor. *Cell* **87,** 175–185 (1996).
73. Ahmed, G. *et al.* Draxin inhibits axonal outgrowth through the netrin receptor DCC. *J. Neurosci.* **31,** 14018–14023 (2011).
74. Rajagopalan, S. *et al.* Neogenin mediates the action of repulsive guidance molecule. *Nature Cell Biol.* **6,** 756–762 (2004).
75. Monnier, P. P. *et al.* RGM is a repulsive guidance molecule for retinal axons. *Nature* **419,** 392–395 (2002).
76. Kajiwara, Y., Buxbaum, J. D. & Grice, D. E. SLITRK1 binds 14-3-3 and regulates neurite outgrowth in a phosphorylation-dependent manner. *Biol. Psychiatry* **66,** 918–925 (2009).
77. Islam, S. M. *et al.* Draxin, a repulsive guidance protein for spinal cord and forebrain commissures. *Science* **323,** 388–393 (2009).
78. Yang, Z. *et al.* NG2 glial cells provide a favorable substrate for growing axons. *J. Neurosci.* **26,** 3829–3839 (2006).
79. Hossain-Ibrahim, M. K., Rezajooi, K., Stallcup, W. B., Lieberman, A. R. & Anderson, P. N. Analysis of axonal regeneration in the central and peripheral nervous systems of the NG2-deficient mouse. *BMC Neurosci.* **8,** 80 (2007).
80. Lu, P., Jones, L. L. & Tuszynski, M. H. Axon regeneration through scars and into sites of chronic spinal cord injury. *Exp. Neurol.* **203,** 8–21 (2007).
81. Busch, S. A. *et al.* Adult NG2 + cells are permissive to neurite outgrowth and stabilize sensory axons during macrophage-induced axonal dieback after spinal cord injury. *J. Neurosci.* **30,** 255–265 (2010).
82. Nakanishi, K. *et al.* Identification of neurite outgrowth-promoting domains of neuroglycan C, a brain-specific chondroitin sulfate proteoglycan, and involvement of phosphatidylinositol 3-kinase and protein kinase C signaling pathways in neuritogenesis. *J. Biol. Chem.* **281,** 24970–24978 (2006).
83. Götz, B. *et al.* Tenascin-C contains distinct adhesive, anti-adhesive, and neurite outgrowth promoting sites for neurons. *J. Cell Biol.* **132,** 681–699 (1996).
84. Andrews, M. R. *et al.* Alpha9 integrin promotes neurite outgrowth on tenascin-C and enhances sensory axon regeneration. *J. Neurosci.* **29,** 5546–5557 (2009).
85. Edwards, T. J. & Hammarlund, M. Syndecan promotes axon regeneration by stabilizing growth cone migration. *Cell Rep.* **8,** 272–283 (2014).
86. Farhy Tselnicker, I., Boisvert, M. M. & Allen, N. J. The role of neuronal versus astrocyte-derived heparan sulfate proteoglycans in brain development and injury. *Biochem. Soc. Trans.* **42,** 1263–1269 (2014).
87. Lu, P., Jones, L. L. & Tuszynski, M. H. BDNF-expressing marrow stromal cells support extensive axonal growth at sites of spinal cord injury. *Exp. Neurol.* **191,** 344–360 (2005).
88. Grill, R., Murai, K., Blesch, A. & Tuszynski, M. H. Cellular delivery of neurotrophin-3 promotes coricospinal axonal growth and partial functional recovery after spinal cord injury. *J. Neurosci.* **17,** 5560–5572 (1997).
89. Blesch, A. & Tuszynski, M. H. Cellular GDNF delivery promotes growth of motor and dorsal column sensory axons after partial and complete spinal cord transections and induces remyelination. *J. Comp. Neurol.* **467,** 403–417 (2003).
90. Blesch, A. *et al.* Leukemia inhibitory factor augments neurotrophin expression and corticospinal axon growth after adult CNS injury. *J. Neurosci.* **19,** 3556–3566 (1999).
91. Cafferty, W. B. *et al.* Leukemia inhibitory factor determines the growth status of injured adult sensory neurons. *J. Neurosci.* **21,** 7161–7170 (2001).
92. Müller, A., Hauk, T. G. & Fischer, D. Astrocyte-derived CNTF switches mature RGCs to a regenerative state following inflammatory stimulation. *Brain* **130,** 3308–3320 (2007).
93. Ozdinler, P. H. & Macklis, J. D. IGF-I specifically enhances axon outgrowth of corticospinal motor neurons. *Nature Neurosci.* **9,** 1371–1381 (2006).

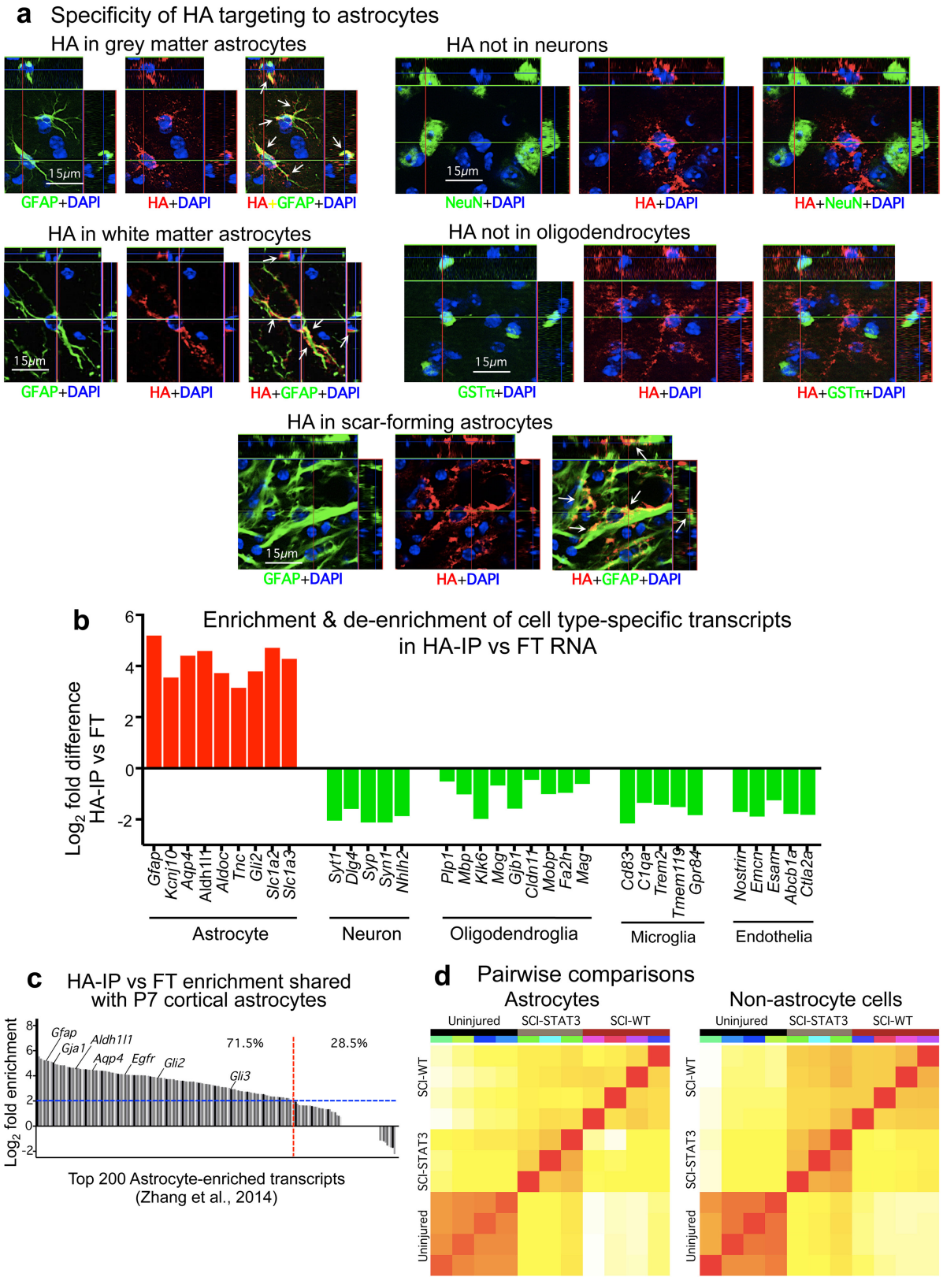
94. Szebenyi, G. *et al.* Fibroblast growth factor-2 promotes axon branching of cortical neurons by influencing morphology and behavior of the primary growth cone. *J. Neurosci.* **21,** 3932–3941 (2001).
95. White, R. E., Yin, F. Q. & Jakeman, L. B. TGF-α increases astrocyte invasion and promotes axonal growth into the lesion following spinal cord injury in mice. *Exp. Neurol.* **214,** 10–24 (2008).
96. Tom, V. J., Doller, C. M., Malouf, A. T. & Silver, J. Astrocyte-associated fibronectin is critical for axonal regeneration in adult white matter. *J. Neurosci.* **24,** 9282–9290 (2004).
97. Qin, J., Liang, J. & Ding, M. Perlecan antagonizes collagen IV and ADAMTS9/GON-1 in restricting the growth of presynaptic boutons. *J. Neurosci.* **34,** 10311–10324 (2014).
98. Hill, J. J., Jin, K., Mao, X. O., Xie, L. & Greenberg, D. A. Intracerebral chondroitinase ABC and heparan sulfate proteoglycan glypican improve outcome from chronic stroke in rats. *Proc. Natl Acad. Sci. USA* **109,** 9155–9160 (2012).
99. Minor, K. *et al.* Decorin promotes robust axon growth on inhibitory CSPGs and myelin via a direct effect on neurons. *Neurobiol. Dis.* **32,** 88–95 (2008).
100. Horie, H. *et al.* Galectin-1 regulates initial axonal growth in peripheral nerves after axotomy. *J. Neurosci.* **19,** 9964–9974 (1999).
101. Walsh, F. S. & Doherty, P. Neural cell adhesion molecules of the immunoglobulin superfamily: role in axon growth and guidance. *Annu. Rev. Cell Dev. Biol.* **13,** 425–456 (1997).
102. Malin, D. *et al.* The extracellular-matrix protein matrilin 2 participates in peripheral nerve regeneration. *J. Cell Sci.* **122,** 995–1004 (2009).
103. Zhang, Y. *et al.* An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *J. Neurosci.* **34,** 11929–11947 (2014).



Extended Data Figure 1 | SCI model schematic, locomotor behavioural effects, and AAV vector targeting specificity and effects. a, Schematic of severe lateral crush SCI at thoracic level T10 that generates a large lesion core (LC) of non-neural tissue surrounded by an astrocytic scar (AS) and completely transects descending and ascending axons. **b**, Open field hindlimb locomotor score at various times after SCI assessed using a 5-point scale where 5 is normal and 0 is no movement of any kind[17]. No significant differences were observed among any of the experimental groups at any time point. $n = 6$ mice at all time points $P > 0.5$ (ANOVA with Newman–Keuls post hoc analysis). WT, wild type. **c**, Horizontal sections through a severe SCI lesion of a representative tdTomato (tdT) reporter mouse[51] injected with an AAV vector with a minimal *Gfap* promoter regulating Cre (AAV2/5-GfaABC1D-Cre) into the lesion at two weeks after SCI and perfused at three weeks. tdTomato labelling demonstrates that this AAV2/5-GfaABC1D-Cre efficiently and specifically targets GFAP-positive astrocytes. In this mouse, the amount AAV2/5-GfaABC1D-Cre injected was intentionally titrated on the basis of previous trial and error to target primarily the astrocytic scar border in an approximately 500 μm zone immediately abutting the SCI lesion core. High-magnification analysis of individual fluorescence channels stained for tdTomato plus various cell markers shows the specificity of Cre activity targeting to cells expressing the astrocyte marker, GFAP, but not to cells expressing either the neuronal marker, NeuN, or the mature oligodendrocyte marker, GSTπ. AAV2/5-GfaABC1D-Cre was prepared using a previously described and well-characterized cloning strategy[53–55]. **d**, Open field hindlimb locomotor scores at various times after SCI. There was no difference in scores of control mice and *loxP*-STOP-*loxP*-DTR (diphtheria toxin receptor) mice that received AAV2/5-GfaABC1D-Cre before injections of diphtheria toxin (DT). Five weeks after DT injections, *loxP*-DTR mice that received AAV2/5- GfaABC1D-Cre exhibited a slightly, but significantly, lower locomotor score. Hindlimb locomotion was assessed using a 5-point scale where 5 is normal and 0 is no movement of any kind[17]. $n = 6$ mice per group; $^*P < 0.05$ versus wild-type (ANOVA with Newman–Keuls). **e**, GFAP immunohistochemistry of a sagittal section after ablation of a chronic astrocytic scar plus adjacent astrocytes. DT was administered to a transgenic mouse expressing DTR targeted selectively to astrocytes around a severe SCI. In this case, the amount of AAV2/5-GfaABC1D-Cre injected was titrated to target not only primarily the astrocytic scar border but also adjacent astrocytes spread over approximately 2 mm on either side of the centre (Cn) of the SCI lesion core (LC). Note the profound degeneration of neural tissue resulting from the selective ablation of the chronic astrocytic scar plus adjacent astrocytes after SCI.

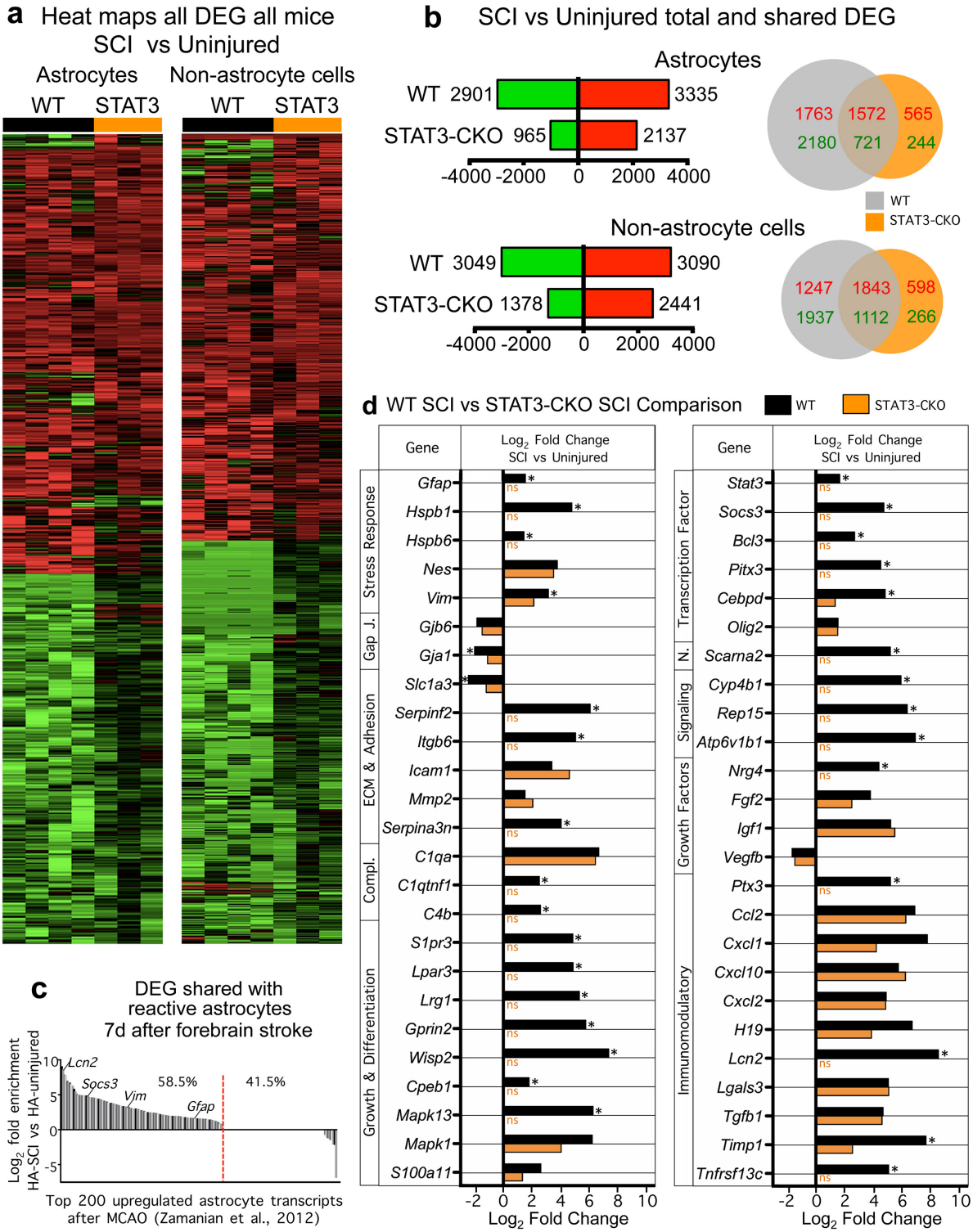


Extended Data Figure 2 | Single-channel CSPG and GFAP immunofluorescence and stained area quantification. a, Individual fluorescence channels of CS56 and GFAP immunohistochemistry from horizontal sections of uninjured mice and at two weeks after severe SCI shown in Fig. 3b. Sections are taken from wild-type (WT) mice and mice with transgenic ablation (TK+GCV) or attenuation (STAT3-CKO) of astrocytic scar formation. **b**, Example of black and white thresholding of single channels of immunofluorescence staining for image analysis to quantify (using NIH Image J software) the amount of CSPG- or GFAP-stained area in different tissue compartments in SCI lesions. Boxes denote areas quantified to obtain values for lesion core (LC) and grey (g) or white (w) matter in astrocytic scar (AS) or equivalent regions in uninjured tissue. Graphs show percentage of areas (means ± s.e.m.) stained for CSPG or GFAP determined using ImageJ. $n = 4$ (wild type mice); $n = 6$ (TK+GCV and STAT3-CKO mice); $\#P < 0.05$ versus uninjured white matter; $*P < 0.05$ versus uninjured grey matter in same experimental group (ANOVA with Newman–Keuls); $\wedge P < 0.05$ versus equivalent anatomical region in wild-type (ANOVA with Newman–Keuls).

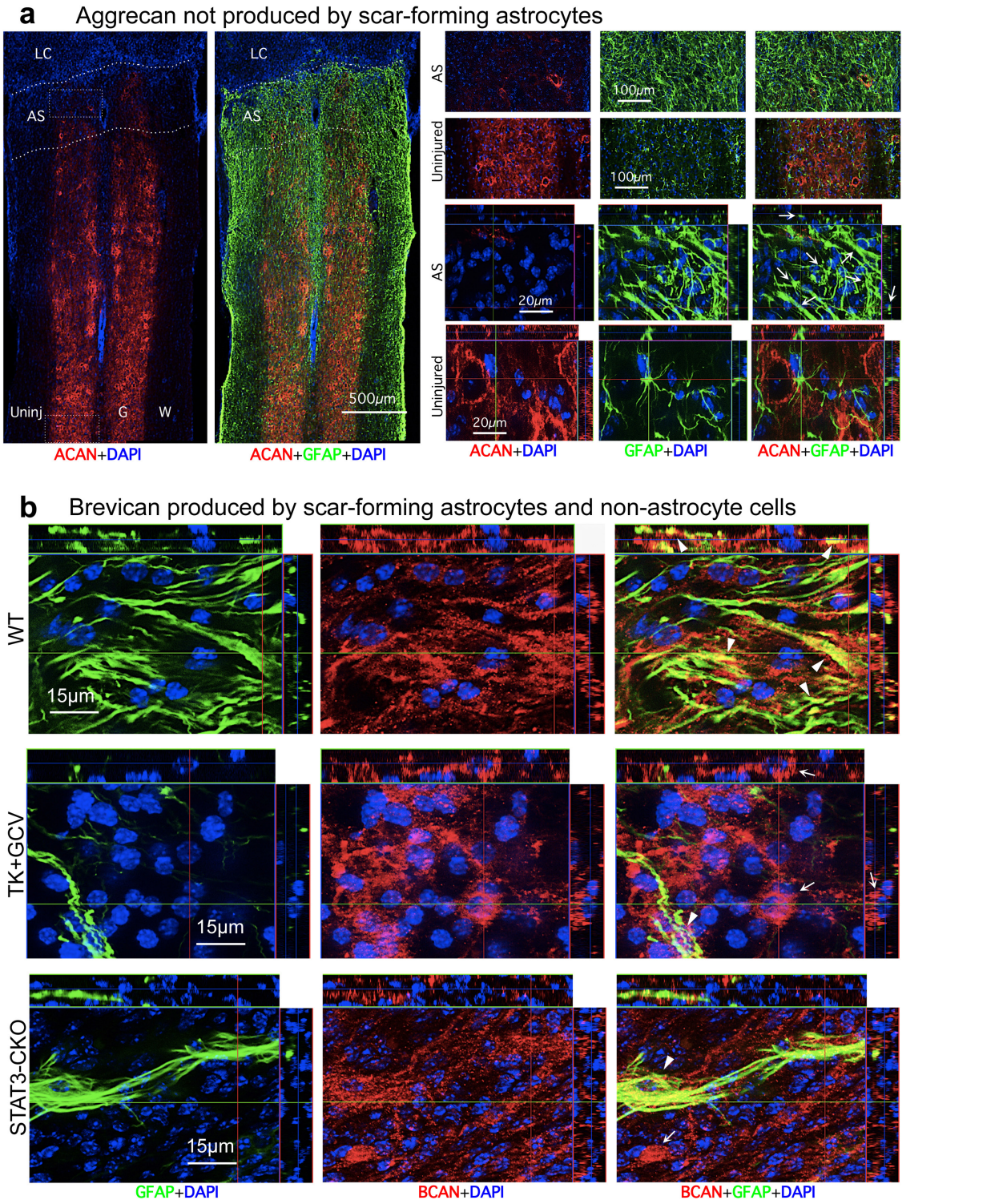


Extended Data Figure 3 | See next page for caption.

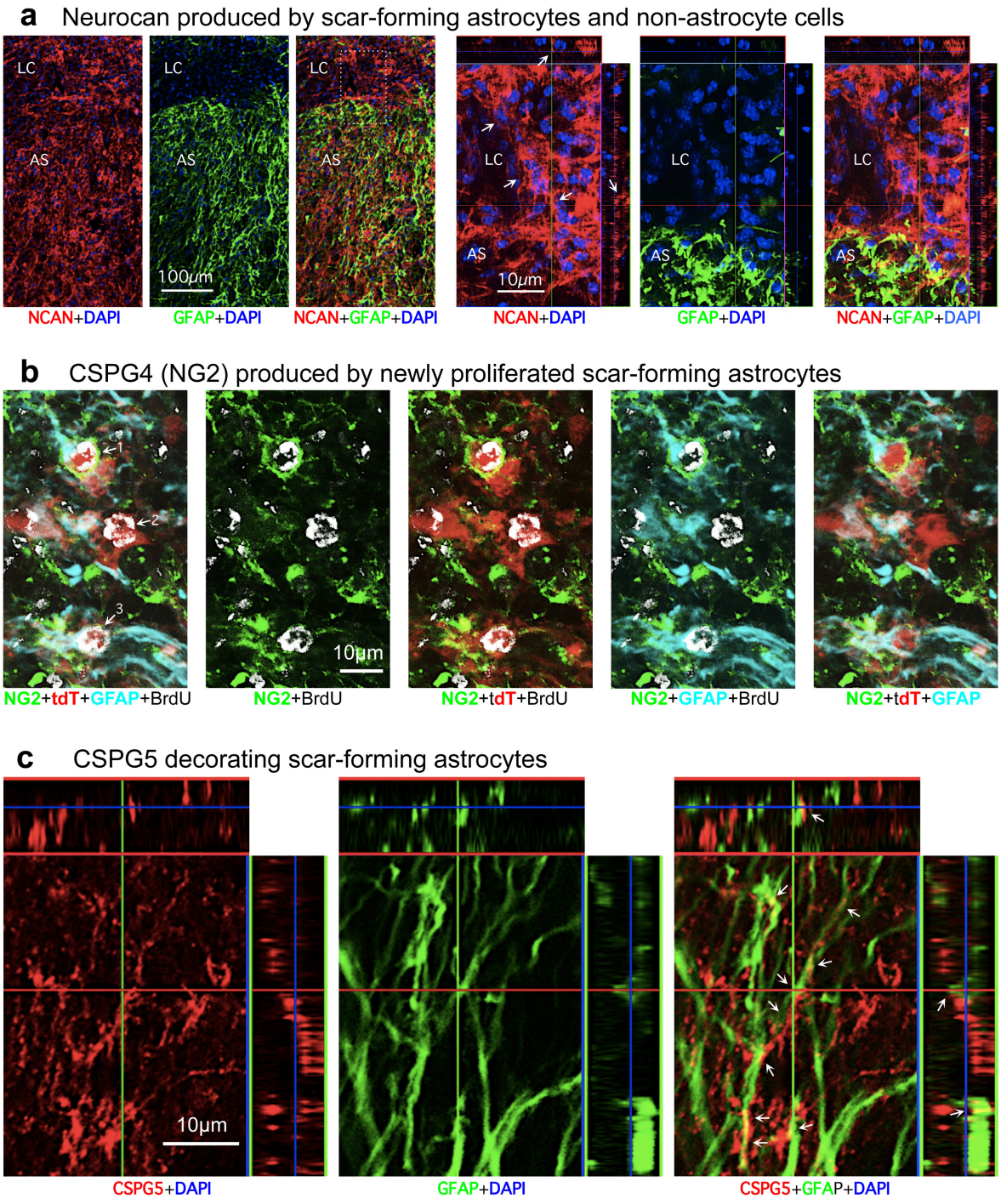
Extended Data Figure 3 | Specificity of haemagglutinin targeting to astrocytes and enrichment of haemagglutinin immunoprecipitation for astrocyte-specific RNA transcripts. a, Individual fluorescence channels of immunohistochemistry for transgenically targeted haemagglutinin (HA) plus various cell markers showing the specificity of HA targeting to cells expressing the astrocyte marker, GFAP, and not to cells expressing either the neuronal marker, NeuN, or the mature oligodendrocyte marker, GSTπ, in uninjured grey and white matter and in astrocytic scars at 2 weeks after SCI. **b**, CNS-cell-type-specific gene transcript enrichment of ribosome-associated mRNA (ramRNA) isolated from wild-type (WT) uninjured spinal cord by HA immunoprecipitation (HA-IP). Differential expression analysis by RNA-seq indicates significant enrichment (red) for astrocyte-specific gene transcripts, and de-enrichment (green) for gene transcripts enriched in other CNS cell types, FDR < 0.1. A $\log_2$ scale is used so that positive and negative differences are directly comparable. The mean numerical enrichment of three quintessential astrocyte genes, *Gfap*, *Aldh1l1* and *Aqp4*, is 25-fold greater in HA samples than in flow-through samples. **c**, Gene transcript enrichment of HA-IP ramRNA relative to P7 mouse primary cortical astrocytes[103]. Of the 200 most highly expressed genes previously described[103] for post-natal mouse cortical astrocytes, 71.5% (red line) are at least fourfold enriched (blue line) in HA-IP ramRNA isolated from uninjured spinal cord relative to flow through RNA from non-astrocyte cells. **d**, Pearson correlation plots of total normalized RNA-seq reads from individual biological replicates for each treatment condition. Correlation colouring indicates little (white) to high (red) similarity. $n = 4$ mice each for uninjured controls and wild-type SCI (SCI-WT); $n = 3$ mice for STAT3-CKO SCI (SCI-STAT3). FDR < 0.1 for differential expression and enrichment analysis. Raw and normalized data have been deposited in the NCBI Gene Expression Omnibus and are accessible through GEO Series accession number GSE76097 and via a searchable, open-access website https://astrocyte.rnaseq.sofroniewlab.neurobio.ucla.edu.



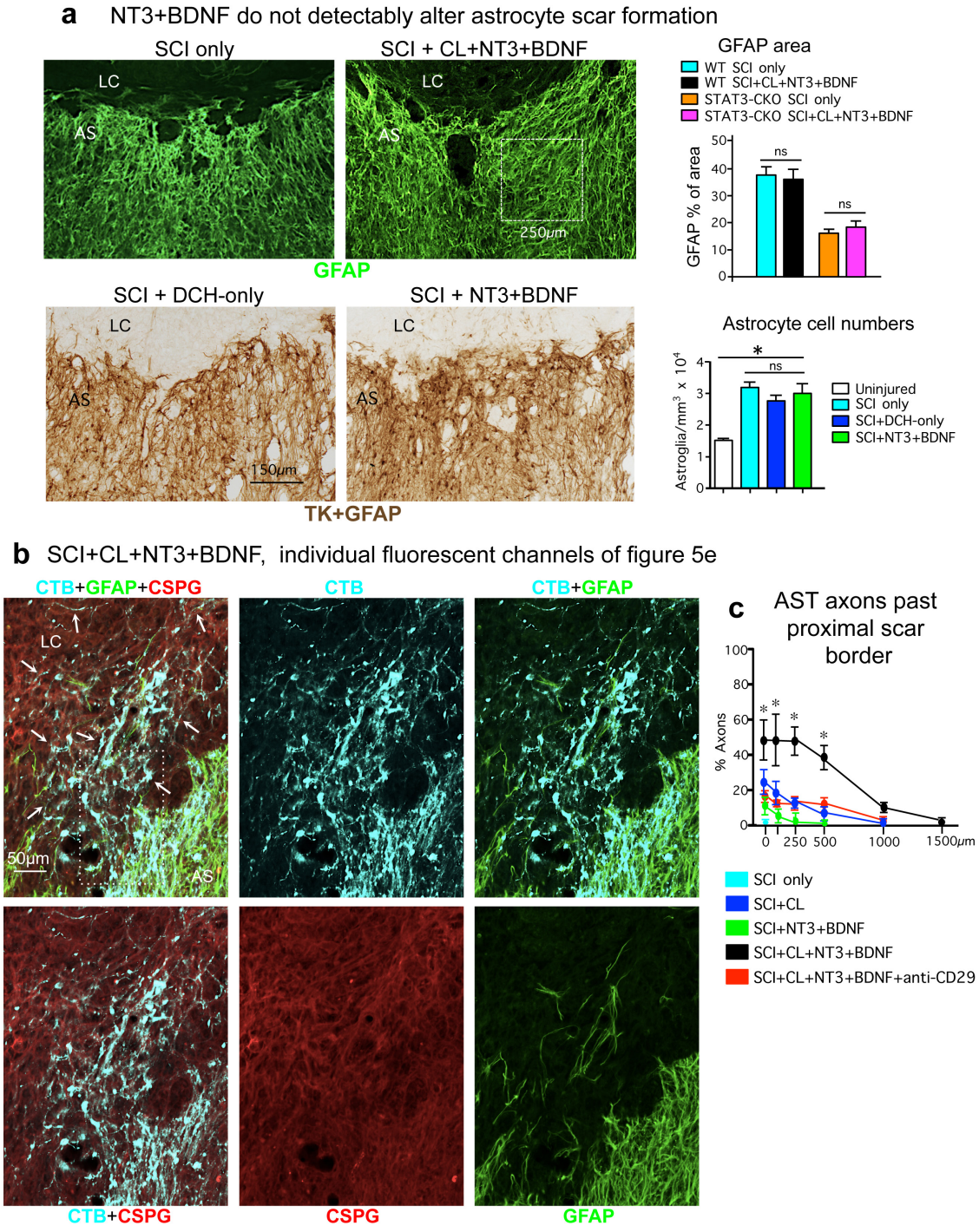
Extended Data Figure 4 | Comparison of genomic data from astrocytes and non-astrocyte cells from WT and STAT3-CKO mice after SCI. **a**, Heat maps depicting all significantly differentially expressed genes (DEG), as determined by RNA-seq, for wild-type (WT) and STAT3-CKO astrocytes and non-astrocytes from independent biological replicates two weeks after SCI relative to uninjured wild-type control. Red upregulated, green downregulated. **b**, Total numbers and Venn diagrams of significant DEGs in wild-type and STAT3-CKO astrocytes and non-astrocytes two weeks after SCI relative to uninjured control. Red and green numerical values indicate significantly upregulated and downregulated genes, respectively. **c**, Comparison of altered gene expression in our SCI-reactive astrocytes and previously reported forebrain stroke-reactive astrocytes[27]. Of the 200 most highly elevated genes in forebrain astrocytes one week following stroke[27], 58.5% (red line) are also significantly elevated in astrocytes after SCI, relative to uninjured. **d**, Comparison of expression by wild-type SCI and STAT3-CKO SCI reactive astrocytes of a selected cross-section of genes that are highly regulated after SCI by wild-type reactive astrocytes. Many of the regulated genes exhibit changes that are expected and implicated in wild-type reactive astrogliosis mechanisms and roles, and some of the changes appear to be newly identified in this context. Note that many of the genes are not regulated or exhibit attenuated changes in STAT3-CKO SCI astrocytes. $n = 4$ mice each for uninjured and wild-type SCI; $n = 3$ mice for STAT3-CKO SCI (SCI-STAT3). FDR < 0.1 for differential expression and enrichment analysis.



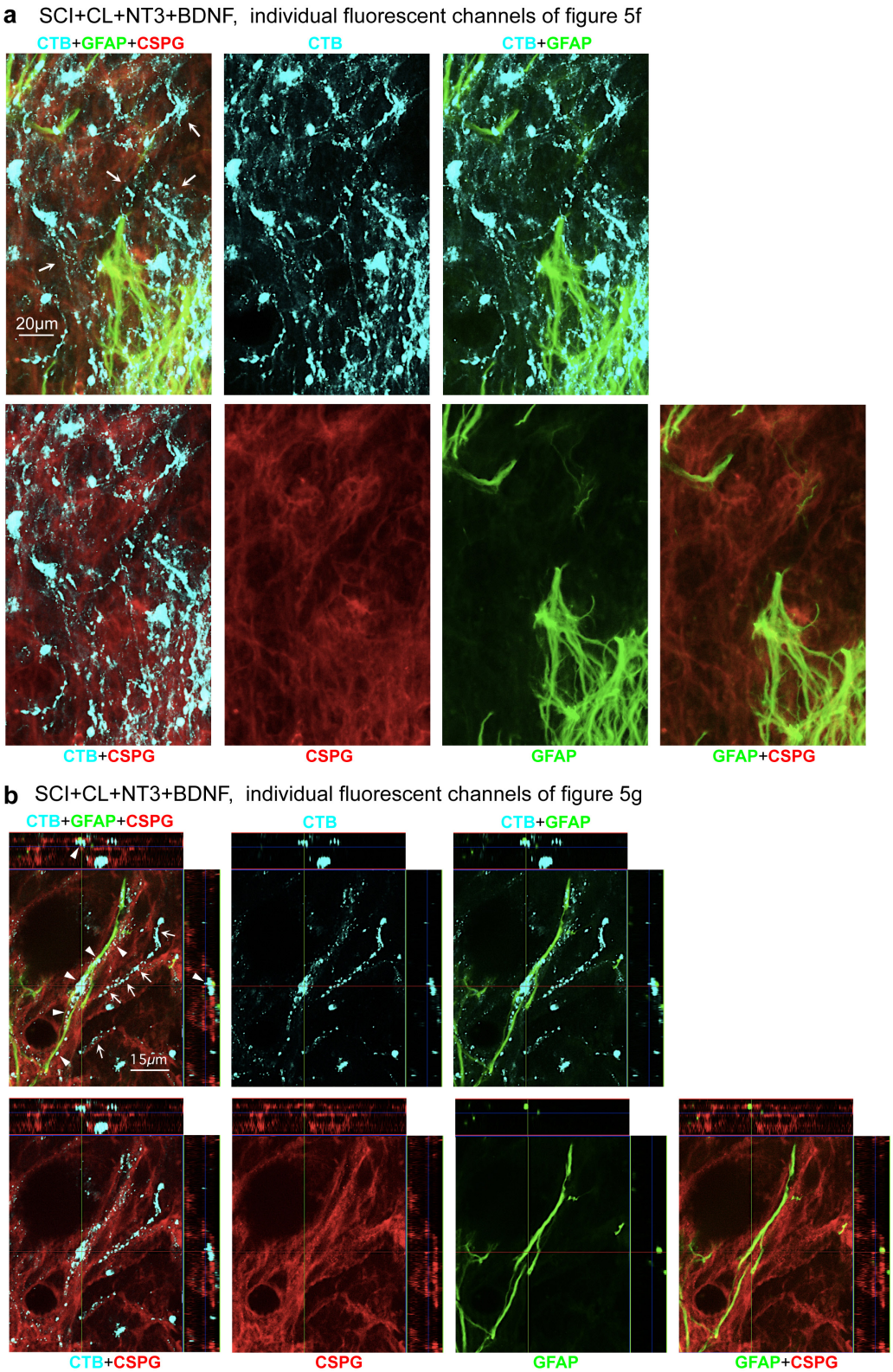
Extended Data Figure 5 | Immunohistochemistry of specific CSPGs. **a**, Absence of aggrecan (ACAN) production by scar-forming astrocytes. Images show individual fluorescence channels of ACAN and GFAP immunohistochemistry from horizontal sections two weeks after severe SCI in a representative wild-type (WT) mouse. Boxes denote areas of astrocytic scar (AS) or uninjured tissue (Uninj) shown at higher magnification. Note that ACAN is: (i) heavily present in the perineuronal nets that surround neurons in uninjured tissue; (ii) almost absent from astrocytic scar and lesion core (LC); and (iii) not detectably produced by newly generated scar-forming astrocytes (arrows). **b**, Brevican (BCAN) production by scar-forming astrocytes and non-astrocyte cells. Images show individual fluorescence channels of BCAN and GFAP immunohistochemistry from horizontal sections two weeks after severe SCI, in wild-type mice and mice with transgenic ablation (TK+GCV) or attenuation (STAT3-CKO) of astrocytic scar formation. Note that BCAN is produced both by GFAP-positive scar-forming astrocytes (arrowheads) and by non-astrocyte cells (arrows).



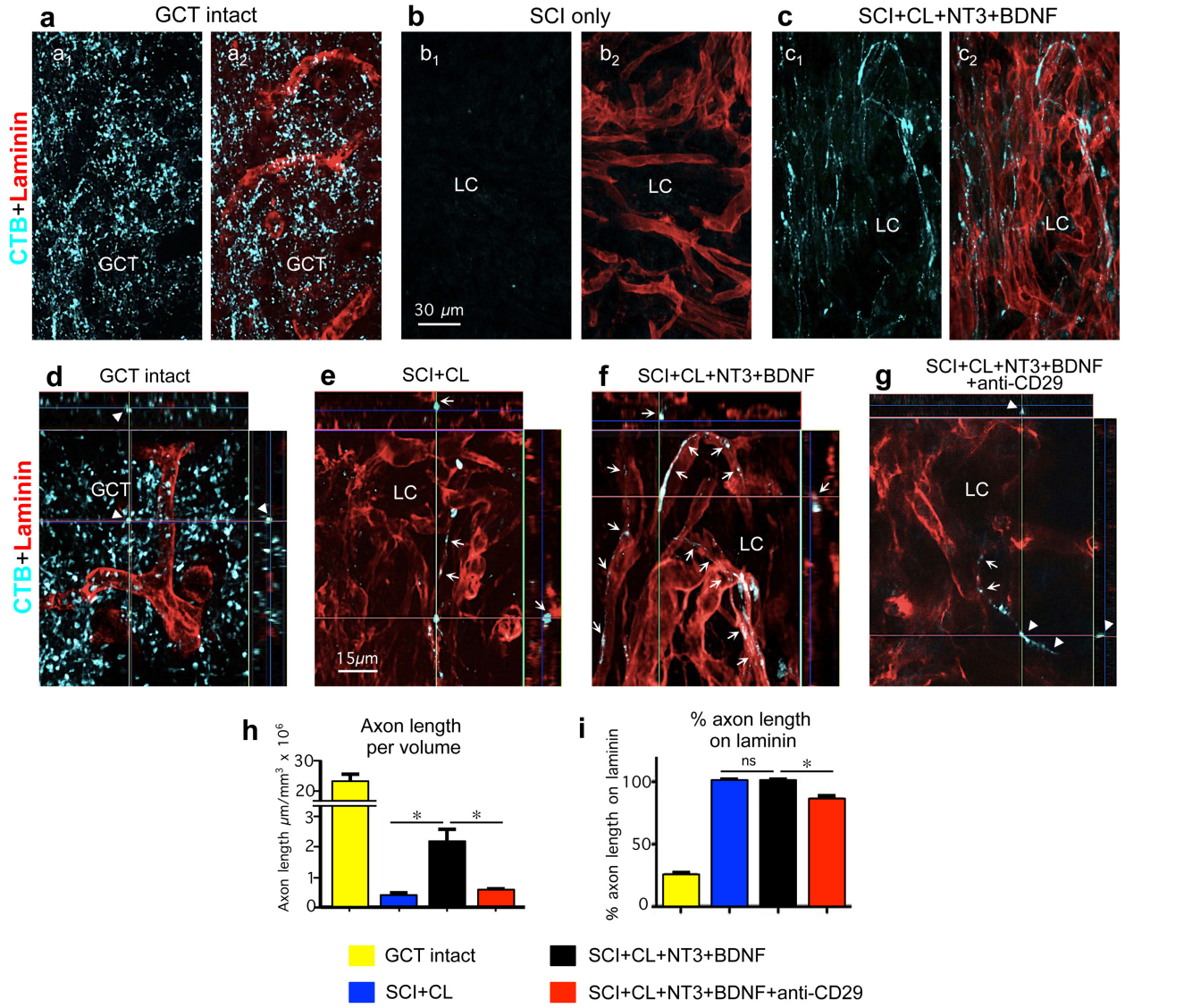
Extended Data Figure 6 | Immunohistochemistry of specific CSPGs. **a**, Neurocan (NCAN) production by scar-forming astrocytes and non-astrocyte cells. Images show individual fluorescence channels of NCAN and GFAP immunohistochemistry from horizontal sections two weeks after severe SCI, in a representative wild-type (WT) mouse. Box denotes area of lesion core (LC) and astrocytic scar (AS) shown at higher magnification. Note that NCAN is produced both by GFAP-positive scar-forming astrocytes and by non-astrocyte cells (arrows) in the lesion core. **b**, NG2 (CSPG4) production by newly proliferated scar-forming astrocytes. Images show individual channels and various combinations of immunofluorescence staining for NG2, GFAP, tdTomato (tdT), BrdU (proliferation marker) and DAPI showing astrocytes in a mature SCI scar. The images are representative of findings from tdTomato-reporter mice[51] injected with AAV2/5-GfaABC1D-Cre vector[53] into multiple sites of the uninjured spinal cord to label mature astrocytes. Three weeks after AAV2/5-GfaABC1D-Cre injection, the mice received a severe SCI and were administered BrdU from days 2–7 after SCI. The mice were perfused after two weeks after SCI. Images comparing individual fluorescence channels show that astrocytes labelled 1 and 3: (i) incorporated BrdU and thus are newly proliferated after SCI; (ii) express the tdTomato reporter; (iii) express GFAP, the prototypical marker of reactive and scar-forming astrocytes; and (iv) express NG2 both intracellularly and along their cell surfaces. In contrast, astrocyte number 2 is also BrdU-labelled and expresses both tdTomato and GFAP, but does not appear to express detectable levels of NG2. **c**, CSPG5 (Neuroglycan C) production by scar-forming astrocytes. Images show individual channels and various combinations of immunofluorescence staining for CSPG5 or GFAP. Note that CSPG5 is present within and along the processes of GFAP-positive scar-forming astrocytes (arrows).



Extended Data Figure 7 | Specificity and effects of treatments to stimulate AST axon regrowth after SCI. a, BDNF and NT3 treatment does not alter the appearance or density of astrocyte scars in wild-type (WT) or STAT3-CKO mice. Images show horizontal sections of mice at two weeks after SCI or after SCI followed by delayed injection of hydrogel only (as a control) or hydrogel releasing NT3 and BDNF. Top images show GFAP immunofluorescence; boxed area denotes size of areas taken from multiple locations in the astrocytic scar (AS) for GFAP area quantification shown in graph. $n = 5$ mice per group; NS, $P > 0.05$ (ANOVA with Newman–Keuls). Bottom images show brightfield immunohistochemistry simultaneously of GFAP+TK to stain both astrocyte cell processes (GFAP) and cell bodies (TK) in mGFAP-TK transgenic mice for quantification of astrocyte cell numbers shown in graph. For these experiments the transgene-derived TK is used as a reporter protein that efficiently labels astrocyte cell bodies and thereby improves cell quantification[18] and the mice were not given GCV. $n = 4$ mice per group; $^*P < 0.05$ versus uninjured (ANOVA with Newman–Keuls); NS, $P > 0.05$ (ANOVA with Newman–Keuls). **b**, AST axon regrowth through scar-forming astrocytes and CSPGs in SCI lesions. Images show individual channels and various combinations of immunofluorescence staining for CTB, GFAP and CS56 to detect total CSPGs from a wild-type mouse after SCI followed by delayed injection of a hydrogel depot releasing NT3 and BDNF, shown as multichannel image in Fig. 5e. Arrows denote robust regrowth of many AST axons along, through and past scar-forming astrocytes into and through the lesion core. Note that the stimulated axons are regrowing through CSPG containing areas in the astrocyte scar and lesion core. Boxed area is shown at higher magnification in Extended Data Fig. 8. **c**, Graph shows numbers of AST axons at various distances past the proximal border of the astrocytic scar under different conditions. $n = 5$ per group; $^*P < 0.001$ significant difference SCI+CL+BDNF+NT3 versus all other groups (ANOVA with post-hoc Newman–Keuls).



Extended Data Figure 8 | AST axon regrowth through scar-forming astrocytes and CSPGs in SCI lesions. a, **b**, Images show individual channels and various combinations of immunofluorescence staining for CTB, GFAP and CS56 to detect total CSPGs from a wild-type (WT) mouse after SCI followed by delayed injection of a hydrogel depot releasing NT3 and BDNF, shown as multichannel images in Fig. 5f, g. Arrows in **a** denote robust regrowth of many AST axons along, through and past scar-forming astrocytes into and through the lesion core; note that the stimulated axons are regrowing through CSPG containing areas in the astrocytic scar and lesion core. **b**, High-magnification orthogonal images of axons in three visual planes. Arrows in **b** denote AST regrowing axons tracking along CSPG-positive and GFAP-negative structures. Arrowheads in **b** denote AST axons tracking along GFAP-positive and CSPG-positive astrocyte processes, passing from one astrocyte process to another.



Extended Data Figure 9 | AST axon regrowth in SCI lesions is dependent on laminin. a–g, Tract tracing of AST axons using CTB and laminin immunohistochemistry. **a–c**, Same fields imaged for CTB alone (a_1–c_1), or CTB plus laminin (a_2–c_2). **a**, **d**, Intact gracile–cuneate tract (GCT). **b**, SCI only. **c**, **f**, SCI plus conditioning lesion (CL) plus hydrogel with growth factors. **e**, SCI plus conditioning lesion. **g**, SCI plus conditioning lesion plus hydrogel with growth factors and anti-CD29. **d–g**, High-magnification orthogonal images of axons in three visual planes. Arrows indicate regrowing axons in direct contact with laminin. Arrowheads indicate axons not in direct contact with laminin in the intact GCT (**d**) or with anti-CD29 treatment (**g**). Note the difference in appearance of axons in the intact gracile–cuneate tract (GCT), which are independent of laminin, compared with regrowing axons in lesion core (LC), which track along laminin. **h**, Axon length per tissue volume (means ± s.e.m.) in intact GCT or in SCI lesions under different conditions. Intact GCT values were not included in ANOVA comparison of other 3 groups. **i**, Percentages (means ± s.e.m.) of AST axon length in direct contact with laminin under different conditions. $n = 5$ mice per group; $^*P < 0.001$ (ANOVA with post-hoc Newman–Keuls); NS, not significant (ANOVA with post-hoc Newman–Keuls).

Extended Data Table 1 | Axon growth inhibitory molecules and axon growth permissive molecules

Extended Data Table 1		Gene abbreviation	Molecule name	References for function
	Axon growth inhibitory molecules	*Acan*	Aggrecan	9,29
		Bcan	Brevican	9,29
		Ncan	Neurocan	9,29,62,63
		Vcan	Versican	9,29
		Ptprz1	Phosphacan	9,63
		Xylt1	Xylosyltransferase 1	64
		Tnr	Tenascin R	65
		Epha4	Ephrin A4	7,66
		Ephb2	Ephrin B2	7,66
		Efnb3	Ephrin B3	7,66
		Ntn1	Netrin 1	7,66,67
		Sema3a	Semaphorin 3a	7,66
		Sema3f	Semaphorin 3f	7,66
		Plxna1	Plexin A1	7,68
		Plxnb1	Plexin B1	7,69
		Nrp1	Neuropilin 1	7,70
		Unc5b	Netrin receptor Unc5b	7,66,71
		Dcc	Deleted in Colorectal Cancer	7,72,73
		Neo1	Neogenin 1	7,74
		Rgma	Repulsive guidance molecule A	7,75
		Rgmb	Repulsive guidance molecule B	7,75
		Slit1	Slit 1	7,66
		Slit2	Slit 2	7,66
		Slitrk1	SLIT and NTRK-like family, member 1	76
		Robo1	Robo 1	7,66
		Robo2	Robo 2	7,66
		Robo3	Robo 3	7,66
		Draxin	Draxin	73,77
	Axon growth permissive molecule	*Cspg4*	NG2	78-81
		Cspg5	Neuroglycan C	82
		Tnc	Tenascin C	83,84
		Sdc1	Syndecan 1	85,86
		Sdc2	Syndecan 2	85,86
		Sdc3	Syndecan 3	85,86
		Sdc4	Syndecan 4	85,86
		Bdnf	Brain derived neurotrophic factor	35,87
		Ntf3	Neurotrophin 3	35,88
		Gdnf	Glial derived neurotrophic factor	89
		Lif	Leukemia inhibitory factor	90,91
		Cntf	Ciliary neurotrophic factor	92
		Igf1	Insulin-like growth factor-1	93
		Fgf2	Fibroblast growth factor 2	94
		Tgfa	Transforming growth factor alpha	95
		Lama1	Laminin A1	36
		Lama2	Laminin A2	36
		Lama4	Laminin A4	36
		Lama5	Laminin A5	36
		Lamb1	Laminin B1	36
		Lamc1	Laminin C1	36
		Col4a1	Collagen 4a1	8
		Fn1	Fibronectin 1	96
		Hspg2	Perlecan	86,97
		Gpc1	Glypican 1	86,98
		Gpc3	Glypican 3	86,98
		Gpc5	Glypican 5	86,98
		Dcn	Decorin	99
		Lgals1	Galectin 1	100
		Ncam1	Neural cell adhesion molecule 1	101
		Matn2	Matrilin	102

This table lists the gene abbreviations and full names and of the 59 axon-growth-modulating molecules whose gene expression levels are presented in Fig. 4. The table also summarizes literature providing evidence for the axon-growth-inhibitory or -permissive effects of each molecule. Refs 62–102 are cited in this table.

ARTICLE

doi:10.1038/nature17644

Modulation of tissue repair by regeneration enhancer elements

Junsu Kang[1], Jianxin Hu[2], Ravi Karra[3], Amy L. Dickson[1], Valerie A. Tornini[1], Gregory Nachtrab[1], Matthew Gemberling[1], Joseph A. Goldman[1], Brian L. Black[2] & Kenneth D. Poss[1]

How tissue regeneration programs are triggered by injury has received limited research attention. Here we investigate the existence of enhancer regulatory elements that are activated in regenerating tissue. Transcriptomic analyses reveal that *leptin b* (*lepb*) is highly induced in regenerating hearts and fins of zebrafish. Epigenetic profiling identified a short DNA sequence element upstream and distal to *lepb* that acquires open chromatin marks during regeneration and enables injury-dependent expression from minimal promoters. This element could activate expression in injured neonatal mouse tissues and was divisible into tissue-specific modules sufficient for expression in regenerating zebrafish fins or hearts. Simple enhancer–effector transgenes employing *lepb*-linked sequences upstream of pro- or anti-regenerative factors controlled the efficacy of regeneration in zebrafish. Our findings provide evidence for 'tissue regeneration enhancer elements' (TREEs) that trigger gene expression in injury sites and can be engineered to modulate the regenerative potential of vertebrate organs.

The capacity for complex tissue regeneration is unevenly distributed among vertebrate tissues and species. Salamanders and zebrafish possess remarkable potential to regenerate tissues like amputated appendages, resected heart muscle, and transected spinal cords[1,2]. Investigations of gene expression and function have generated molecular models for regeneration in multiple contexts, yet there is a gap to be filled in our understanding of the regulatory events that activate tissue regeneration programs[1–5].

Recent genome-wide chromatin analyses suggest that gene regulatory elements comprise a substantial portion of genomic sequence. Of these elements, distal-acting regulatory sequences, or enhancers, represent the most abundant class[6,7]. Enhancers can direct expression of their target genes and have been predominantly examined as a means for stage- and tissue-specific regulation during embryonic development[8,9]. Studies have also implicated enhancers in disease and as targets during evolution[10–15]. Such findings raise the possibility that enhancer elements may also exist that engage with transcription factors in response to tissue damage to regulate genetic programs for regeneration. The identification of such elements could potentially inspire new solutions for manipulating regenerative events.

leptin b induction during fin and heart regeneration

To identify genes that are induced during tissue regeneration, we collected RNA from uninjured and regenerating tissues of adult zebrafish and sequenced transcriptomes. Our analyses identified 2,408 genes with significantly higher expression in tail fins at 4 days post-amputation (dpa), and 859 genes with significantly higher expression in cardiac ventricles 7 days after induced genetic ablation of half of all cardiomyocytes (Extended Data Fig. 1a and Supplementary Tables 1 and 2). In total, 360 genes were induced twofold or greater in both tissues compared to uninjured tissues (Extended Data Fig. 1a). Among these genes, 69 were present at low levels in uninjured fins and highly induced during regeneration (Supplementary Information). The gene *leptin b (lepb)*, one of two zebrafish paralogues related to mammalian leptin, a secreted regulator of energy homeostasis[16], had the highest relative change during fin regeneration of genes in this group (130-fold; Fig. 1c, Extended Data Fig. 2, and Supplementary Information). *lepb* transcripts were rare or undetectable in uninjured fins by semi-quantitative or quantitative RT–PCR (qPCR) or *in situ* hybridization (ISH), but induced in the regeneration blastema by 1 dpa (Extended Data Fig. 1b–d). Upon local injury of the cardiac ventricle by partial resection, *lepb* expression was induced in the endocardium, the endothelial lining of inner myofibres that has been implicated in regenerative events (Extended Data Fig. 1b, c, e)[17,18].

To capture the regulatory elements responsible for *lepb* induction, we replaced the first exon of *lepb* with an eGFP reporter transgene within a 150 kb BAC containing 105 kb of DNA sequence upstream of the start codon (Fig. 1d). Transgenic *lepb:eGFP* larvae had little or no detectable eGFP as viewed under a stereofluorescence microscope, and no fluorescence was detectable in fins or hearts throughout life (Fig. 1e and Extended Data Fig. 1h, i, l, m). Upon fin amputation, *lepb*:eGFP fluorescence was sharply induced in regenerating structures, where fluorescence localized to blastemal mesenchyme (Fig. 1e and Extended Data Fig. 1j, k). *lepb*:eGFP was also induced in wounds of resected ventricles, as well as in atrial tissue distant from the site of injury (Fig. 1g), a signature observed with other injury-induced markers[17,18]. While sparse *lepb*:eGFP could be detected in epicardial tissue at 1 day post-resection (dpa; data not shown), cardiac *lepb*:eGFP fluorescence was predominantly endocardial by 3 dpa (Fig. 1g, h). Thus, sequences within a ~150 kb genomic region surrounding *lepb* direct regeneration-dependent expression in fin and cardiac tissues.

lepb-linked enhancer directs expression after injury

Enhancers are identifiable as areas of open chromatin, bound by transcription factors and occupied by histones possessing various modifications, such as acetylated lysine 27 of histone H3 (H3K27ac)[19,20]. To define areas of open chromatin, we assayed genomic regions surrounding *lepb* for H3K27ac marks by ChIP-seq in samples of uninjured and regenerating hearts. Two regions within the *lepb* BAC, located 7 kb and 3 kb upstream of the *lepb* start codon, displayed enrichment with H3K27ac marks in regenerating, but not uninjured, samples (Fig. 2a and Extended Data Fig. 3a, b). To examine if either of these distal

[1]Department of Cell Biology, Duke University Medical Center, Durham, North Carolina 27710, USA. [2]Cardiovascular Research Institute, University of California, San Francisco, San Francisco, California 94143, USA. [3]Department of Medicine, Duke University Medical Center, Durham, North Carolina 27710, USA.

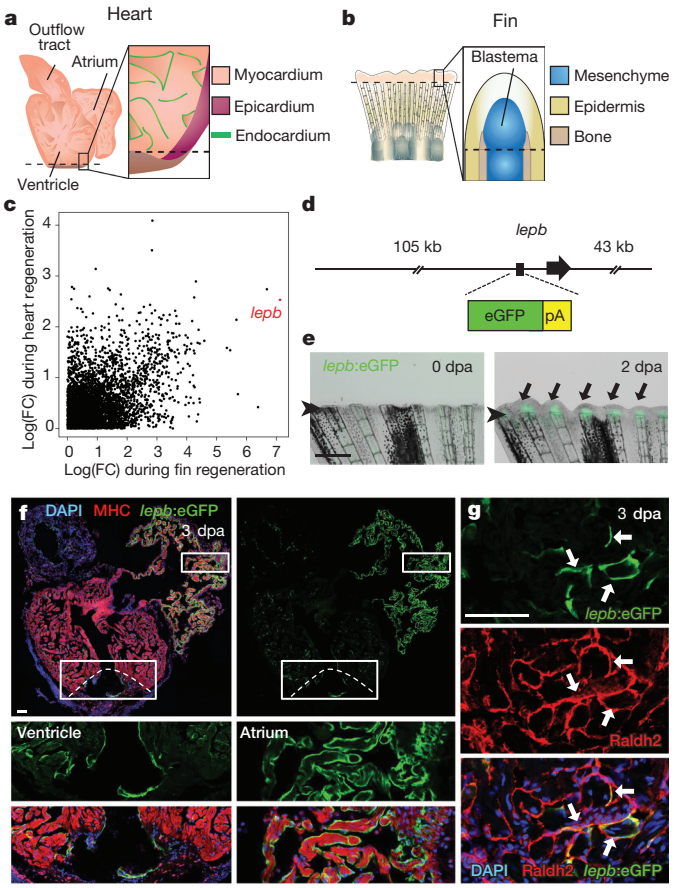


Figure 1 | Activation of *lepb* regulatory sequences during tissue regeneration. a, **b**, Regenerating heart (**a**) and fin (**b**) tissues. **c**, Genes with increased transcript levels in regenerating fins and/or hearts. *lepb* is in red. FC, fold-change. **d**, *lepb:eGFP* BAC transgenic construct, with the first exon replaced by *eGFP*. **e**, *lepb*:eGFP fluorescence (arrows) is detected in fins regenerating after amputation. dpa: days post-amputation. Arrowheads, amputation plane. **f**, **g**, *lepb*:eGFP fluorescence is undetectable in uninjured hearts (see Extended Data Fig. 1), but induced in regenerating hearts by 3 dpa. *lepb*:eGFP fluorescence (arrows in **g**) does not co-localize with MHC^{+} cardiomyocytes (**f**), but co-localizes with Raldh2^{+} endocardial cells (**g**). Antibodies detected eGFP, MHC and Raldh2 in **f**, **g**. $n = 8$; all animals displayed a similar expression pattern. Scale bars represent 500 μm (**e**); 50 μm (**f**, **g**).

regions exhibited enhancer activity, we established several transgenic lines containing 2 kb, 6 kb, and 7 kb upstream sequences of *lepb* fused to an eGFP reporter gene (referred to hereafter as *P2:eGFP*, *P6:eGFP*, and *P7:eGFP*) (Fig. 2b and Extended Data Fig. 3c). Upon fin amputation, only *P7:eGFP* animals, with regulatory sequences encompassing the distal H3K27ac-rich area in the transgene, displayed strong blastemal expression that was comparable to *lepb:eGFP* BAC transgenic animals (Fig. 2c and Extended Data Fig. 3d; *P6:eGFP* fins showed expression below the amputation site). Similarly, whereas *P2:eGFP* and *P6:eGFP* animals occasionally displayed induced fluorescence in myocardium and epicardium after cardiac injury, only injured *P7:eGFP* hearts displayed strong endocardial fluorescence (Fig. 2c and Extended Data Fig. 3f). Thus, a short DNA element located 7 kb upstream of the *lepb* coding sequence is important for directing gene expression in regenerating adult tissues.

We next examined whether an isolated 1.3 kb sequence that corresponded to the H3K27ac-rich region could activate gene expression when fused to *P2*, the presumed *lepb* promoter (Fig. 2b and Extended Data Fig. 3c). Although reporter eGFP fluorescence was not evident in uninjured adult fins or hearts of transgenic fish containing this *lepb*-linked distal element, fin amputation and ventricular resection activated eGFP fluorescence in blastemal and endocardial cells, respectively, in a similar manner to the *lepb* BAC sequences (Fig. 2c and Extended Data Fig. 3d–f). From a genome-wide H3K27ac survey, we also identified many 1–2 kb intergenic regions at other genomic loci that acquired H3K27ac marks during regeneration. We assessed sequence conservation and examined potential enhancer activity by transient transgenic reporter assays using several regions, some of which enabled expression from a minimal *lepb* promoter after injury (Extended Data Fig. 4a–d and Supplementary Information). To further validate the *lepb*-linked element, we examined its ability to influence the cell type-specific promoters *cmlc2* (cardiomyocytes) and *α-cry* (lens) in stable transgenic reporter lines. Robust, regeneration-dependent eGFP fluorescence was evident in fins and hearts of transgenic animals harbouring either the *cmlc2* or *α-cry* promoters (Extended Data Figs 5 and 9a, b, d, g). Thus, a small intergenic element we now refer to as *lepb*-linked enhancer, or *LEN*, can direct regeneration-activated gene expression from multiple promoters.

LEN-associated expression in injured mouse tissues

Analysis of regions upstream of leptin genes in murine and human genomes revealed limited primary sequence conservation of *LEN*

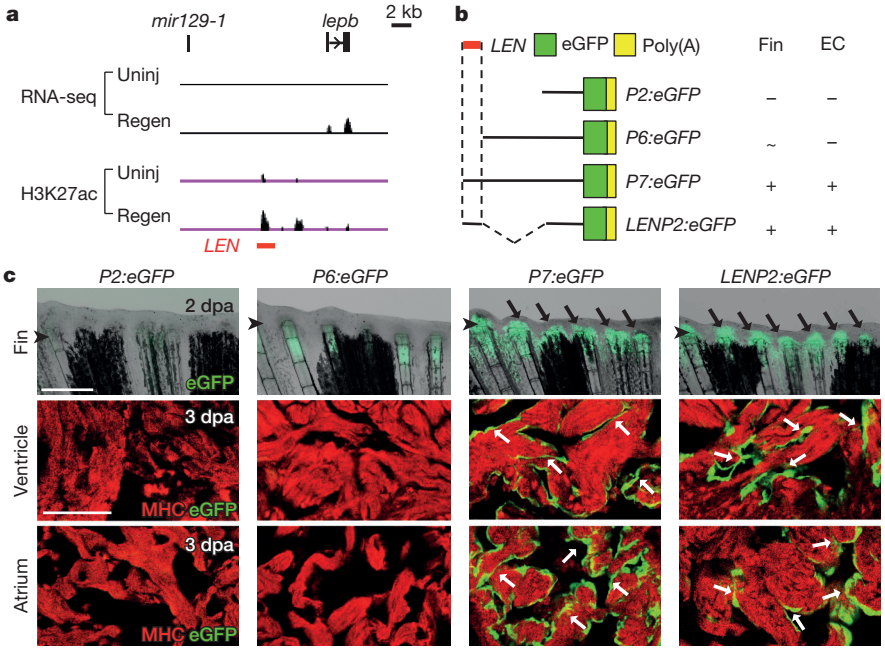


Figure 2 | A DNA element upstream of *lepb* directs regeneration-dependent gene expression. a, Genomic DNA regions surrounding *lepb*, indicating RNA-seq and H3K27ac profiles from uninjured and regenerating hearts. Red bar, distal *lepb*-linked element enriched with H3K27ac marks (*LEN*). **b**, Transgene constructs examined for regeneration-dependent expression in fin or heart. EC, endocardial cells. **c**, Top: images of 2 dpa regenerating fins from transgenic reporter lines. Arrowhead, amputation plane. Arrows, blastemal eGFP. Middle: section images of resected ventricular region at 3 dpa. Bottom: atrial tissue distant from injury site. At least 5 fish from each transgenic line were examined, and all animals displayed a similar expression pattern. Arrows, endocardial eGFP. Scale bars represent 500 μm (top); 50 μm (middle).

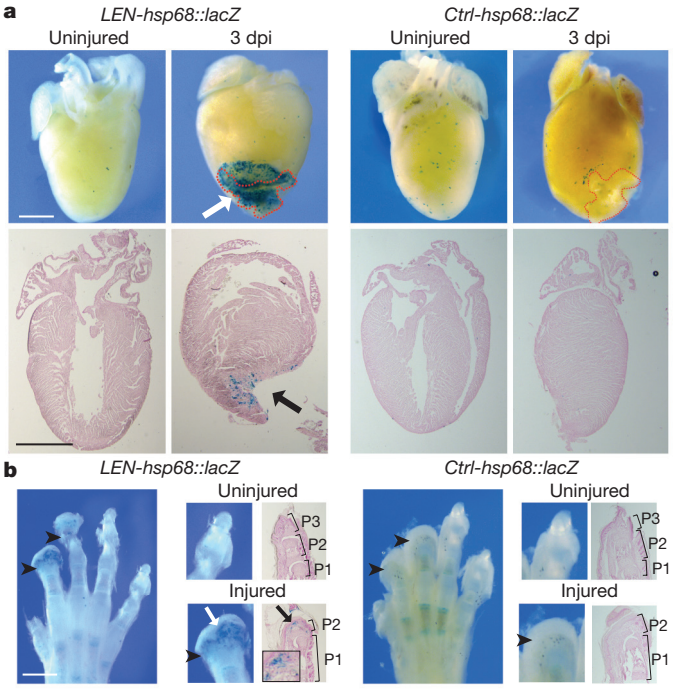


Figure 3 | *LEN* activity in neonatal mice. a, Whole-mount (top) and section (bottom) images of X-gal stained hearts of *LEN-hsp68::lacZ* and *Ctrl-hsp68::lacZ* (control) lines, with clear staining in partially resected hearts of *LEN-hsp68::lacZ* mice (arrows) but not controls. $n = 5, 5, 6$, and 4 for uninjured *LEN-hsp68::lacZ*, 3 days post-injury (dpi) *LEN-hsp68::lacZ*, uninjured control, and 3 dpi control hearts, respectively. Six sham-operated hearts showed minimal staining (see Extended Data Fig. 6). Dashed red lines indicate injury area, positioned facing the front. Arrows, injury-dependent β-galactosidase expression. **b**, Whole-mount (left) and section (right) images of X-gal-stained digits from these lines, with X-gal staining detectable in amputated, but not uninjured, digits of *LEN-hsp68::lacZ* mice. $n = 14(7)$ and 12(6) for *LEN-hsp68::lacZ* and control digits (animals), respectively. Injuries were performed in neonatal mice on postnatal day 1 and assessed for expression on postnatal day 4. Arrowheads, injury planes. Arrows, injury-dependent β-galactosidase expression. P1, P2, P3, proximal, middle, and distal phalange, respectively. Scale bars represent 1 mm.

(Extended Data Fig. 4e). This sequence divergence likely reflects rapid evolution of enhancers, reported in previous studies[21,22]. To examine whether zebrafish *LEN* has activity in mammalian injury contexts, we fused it upstream of a construct containing a murine minimal *hsp68* promoter and a *lacZ* reporter gene. We generated two stable lines, one of which displayed vascular endothelial X-gal staining in uninjured neonatal hearts and paws (Extended Data Fig. 6b). A second line had a small number of X-gal-positive cells in uninjured neonatal tissues and was selected for injury studies (*LEN-hsp68::lacZ*) (Fig. 3a, b). Neonatal digit tips amputated at P2 phalanges do not regenerate lost structures effectively[23], whereas injured neonatal ventricles display a regenerative response[24]. Strikingly, amputated digit tips and damaged ventricles of all injured postnatal day 1 *LEN-hsp68::lacZ* neonates showed conspicuous X-gal staining in wounds 3 days after surgeries. A control transgenic line with an unrelated enhancer fragment also exhibited low basal expression in uninjured neonatal tissues, but unlike *LEN-hsp68::lacZ* animals, showed no detectable activation of the *lacZ* reporter upon injury to the digits or ventricle (Fig. 3a, b and Extended Data Fig. 6a). While future tests of *LEN* activity using a panel of promoters and transgene integration sites will be important, overall, these results suggest that zebrafish *LEN* sequences can interact with mammalian transcriptional machinery to enable injury-induced expression in mice.

LEN is separable into tissue-specific modules

To identify minimal sequences responsible for the activity of *LEN*, we tested the ability of various fragments to direct regeneration-activated expression. We found that more distal *LEN* fragments composed of approximate nucleotides 1–850, 450–1000, 450–850 or 660–850 could each drive eGFP expression from the *lepb* 2 kb promoter during fin regeneration (Fig. 4a, b and Extended Data Fig. 7). *LEN* fragments generated from the distal 1 kb portion also directed eGFP expression during fin regeneration when paired with the *cmlc2* promoter (Extended Data Figs 5 and 9a, b). *LEN* fragments 1–850 and 450–1000 did not direct detectable eGFP expression during fin regeneration from the α-*cry* promoter in our experiments (Extended Data Fig. 5 and 9d–f), suggesting a repressive motif in α-*cry* upstream sequences. Intriguingly, none of these fragments directed endocardial expression after cardiac injury, although eGFP fluorescence was occasionally observed sparsely in epicardial cells or cardiomyocytes (Extended Data Fig. 8). Conversely, more proximal *LEN* fragments comprising approximate nucleotides 830–1350 or 1000–1350 directed endocardial expression during heart regeneration, but did not activate eGFP fluorescence in regenerating fins (Fig. 4a, b and Extended Data Figs 7 and 8). These proximal *LEN* fragments also could direct regeneration-associated expression in endocardial cells from *cmlc2* and α-*cry* promoters

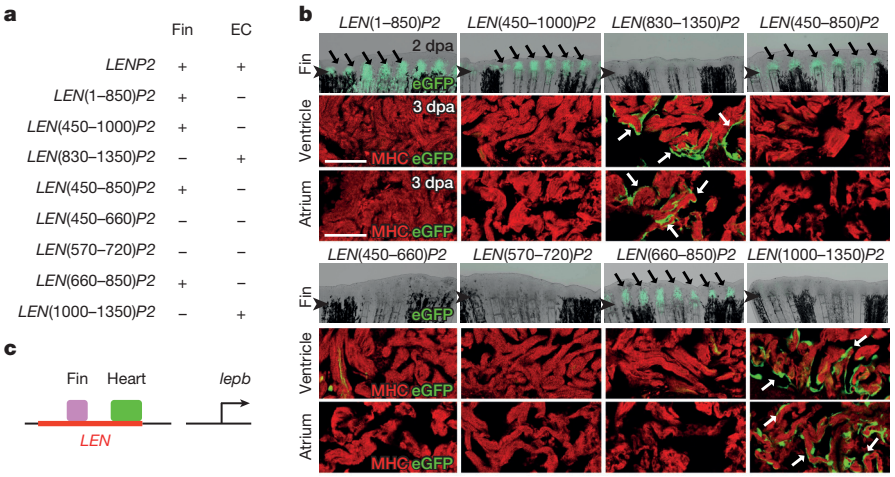


Figure 4 | *LEN* is separable into tissue-specific elements. a, Transgene constructs to examine enhancer activation in regenerating fin or cardiac tissue. EC, endocardial cells. **b**, Regenerating fins (top) and sections of cardiac tissue from transgenic lines in **a**. Middle, resected ventricle region. Bottom, atrial tissue distant from injury site. At least 5 fish from each transgenic line were examined, and all animals displayed a similar expression pattern. Arrowheads, amputation plane. Arrows, blastemal (fin) or endocardial (heart) eGFP. **c**, Cartoon indicating separable tissue-specific regeneration modules in *LEN*. Scale bars, 50 μm (**b**).

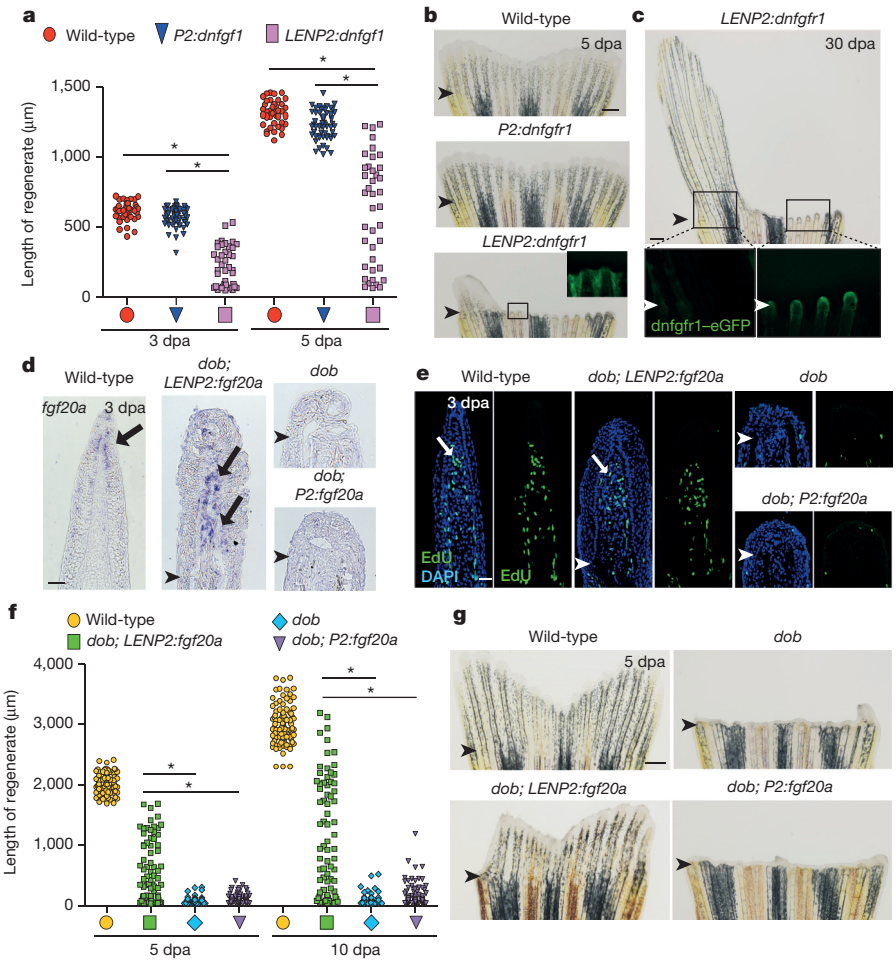


Figure 5 | *LEN* controls fin regeneration when paired with Fgf effectors. **a**, Quantification of third and fourth ray lengths from each lobe at 3 and 5 dpa. *$P < 0.01$, one-way ANOVA; $n = 40$ (10), 56 (14), and 40 (10) for wild-type, *P2:dnfgfr1,* and *LENP2:dnfgfr1* fin rays (number of animals indicated within the brackets), respectively. **b**, Representative images of 5 dpa fin regenerates that were used for quantification of regenerate lengths in **a**. Bottom, inset indicates dnfgfr1–eGFP fluorescence from boxed area. **c**, Images of 30 dpa *LENP2:dnfgfr1* fin regenerate. eGFP fluorescence from boxed areas (left and right, with high-magnification images below), maintained in impaired rays (right). **d**, Section ISH for *fgf20a* expression (arrows) in wild-type, *dob*; *LENP2:fgf20a*, *dob*, and *dob; P2:fgf20a* fin regenerates at 3 dpa. **e**, 3 dpa fin regenerates from animals in **d**, stained for EdU incorporation (green) and nuclei (DAPI, blue), indicating extensive blastemal proliferation (arrows) in wild-type and *dob; LENP2:fgf20a* regenerates. Fins were collected 60 min after EdU injection. **f**, Quantification of third and fourth ray lengths from each lobe at 5 and 10 dpa. *$P < 0.01$, one-way ANOVA; $n = 100$ (25), 72 (18), 56 (14), and 100 (25) for wild-type, *dob; LENP2:fgf20a, dob,* and *dob; P2:fgf20a* fin rays (animals) at 5 dpa, respectively; $n = 98$ (25), 72 (18), 56 (14), and 96 (24) at 10 dpa, respectively. **g**, Representative images of 5 dpa fin regenerates that were used for quantification of regenerate lengths in **f**. The *LENP2:fgf20a* transgene rescues fin regeneration in *dob* animals, shown with controls at 5 dpa. Arrowheads in **b–e, g**, amputation planes. Scale bars represent 500 μm (**b**, **c**, **g**); 20 μm (**d**, **e**).

(Extended Data Fig. 9c, h). Thus, our analyses suggested the presence of two separate, tissue-specific enhancer modules (Fig. 4c).

We analysed sequences of the minimal 190 nucleotide (nt) (fin) and 316 nt (heart) elements, and identified distinct sets of predicted transcription factor binding motifs. *LEN*(663–854) contains predicted AP-1, Sox, forkhead, and ETS binding sites, and we confirmed by transgenic reporter assays that a predicted AP-1 binding site at *LEN*(776–782) is necessary to direct expression in regenerating fins (Extended Data Fig. 9i, j). *LEN*(1034–1350) contains predicted NFAT, GATA, forkhead, and ETS binding sites, motifs associated with expression in endothelial cells[25,26] (Extended Data Fig. 9i). In total, our findings indicate a composite arrangement of regulatory elements with distinct tissue preferences within the *LEN* regeneration enhancer.

LEN element constructs control regenerative capacity

Recent studies have described new enhancer-target gene pairings caused by chromosomal rearrangements that underlie genetic diseases like cancer and neurological disorders[10,12,15]. To examine a parallel idea for experimentally guiding tissue regeneration, we designed transgenic constructs positioning *LEN* and the minimal *lepb* promoter upstream of pro- or anti-regenerative factors. A possible outcome is that *LEN* would limit embryonic expression of potent developmental influences to permit maturation from the one-cell stage to adulthood, but also trigger and sustain expression of these influences upon tissue damage.

To create enhancer-effector transgenes, we took advantage of the dependency of fin regeneration on signalling by fibroblast growth factors (Fgfs)[4,27]. We first positioned *LEN* upstream of a cDNA encoding a dominant-negative form of *fgfr1* (*dnfgfr1*)—a potent inhibitor of embryonic development[27,28]—and injected this construct into wild-type embryos. We established stable lines of zebrafish harbouring either *P2:dnfgfr1* or *LENP2:dnfgfr1*, demonstrating that *dnfgfr1* expression was limited to developmentally insignificant levels. Adult *P2:dnfgfr1* fins displayed no detectable *dnfgfr1* induction after amputation and regenerated normally. By contrast, injury to *LENP2:dnfgfr1* animals induced strong expression of *dnfgfr1* (detectable by dnfgfr1–eGFP fusion protein fluorescence) that was restricted to the amputation plane. Moreover, these animals displayed conspicuous defects or outright failures in fin regeneration (Fig. 5a, b). In some cases, fin rays failed to regenerate even by 30 dpa and maintained *dnfgfr1* expression

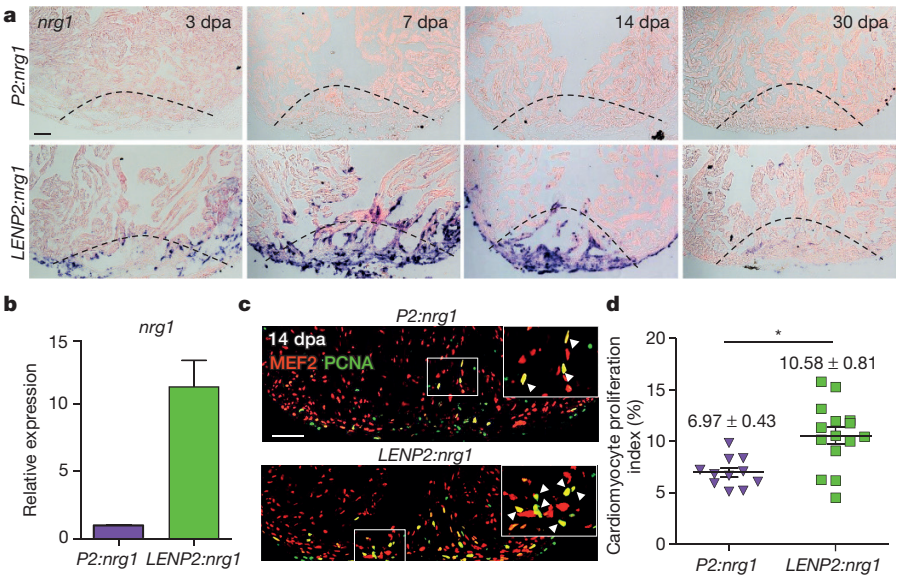


Figure 6 | Enhancer-directed *nrg1* expression boosts cardiomyocyte proliferation. a, Representative images of section ISH for *nrg1* in *P2:nrg1* (top) and *LENP2:nrg1* (bottom) ventricles, at several times post-resection. *P2:nrg1: n* = 4, 8, 7, and 3 for 3, 7, 14, and 30 dpa, respectively. *LENP2:nrg1: n* = 4, 8, 8, and 4 for 3, 7, 14, and 30 dpa, respectively. Dashed lines, approximate resection planes. *nrg1* (violet) is strongly induced in endocardial and epicardial cells in *LENP2:nrg1* ventricular injuries. **b**, qPCR analysis of *nrg1* in whole *P2:nrg1* or *LENP2:nrg1* cardiac ventricles at 3 dpa. Data represent mean ± standard error. *n* = 3. **c**, Section images of 14 dpa regenerating ventricular apices from *P2:nrg1* (top) and *LENP2:nrg1* (bottom) animals, stained for cardiomyocyte nuclei (MEF2; red) and the proliferation marker PCNA (green). Insets indicate a high-magnification view of regenerating area. Arrowheads, $MEF2^{+}PCNA^{+}$ cardiomyocytes. **d**, Quantified cardiomyocyte proliferation indices in injury sites in experiments from **c**. Numbers indicate mean ± standard error. $^{*}P < 0.01$, Mann–Whitney rank sum test; *n* = 11 (*P2:nrg1*) and 15 (*LENP2:nrg1*). Scale bars represent 50 μm (**a**, **c**).

in ray stumps, indicating persistent activation of *LEN* in the setting of regenerative failure (Fig. 5c and Extended Data Fig. 10b).

We complemented these experiments with a gain-of-function approach, based on the discovery that mutations in the *fgf20a* ligand gene, *devoid of blastema* (*dob*), arrest fin regeneration[4]. We positioned *LEN* and the minimal *lepb* promoter upstream of a *fgf20a* cDNA and injected this construct into one-cell *dob* embryos. We generated stable lines of control *dob*; *P2:fgf20a* and *dob; LENP2:fgf20a* animals, indicating that these constructs restricted ectopic *fgf20a* expression during embryonic development. Upon amputation of adult tail fins, *dob*; *P2:fgf20a* animals induced no additional detectable *fgf20a* and displayed regenerative blocks comparable to *dob* animals (Fig. 5d, f, g). By contrast, *LENP2* sequences directed broad expression of *fgf20a* in mesenchymal cells upon fin amputation (Fig. 5d, f, g). Remarkably, blastemal cell proliferation was stimulated in amputated *dob*; *LENP2:fgf20a* fins, and these animals regenerated patterned structures that were often of normal length (Fig. 5e–g). In some cases, the lobed pattern of the tail fin was restored, and in no cases were there uncontrolled growth phenotypes (Fig. 5g).

Targeted cardiomyocyte proliferation by *LEN*

Heart regeneration occurs through injury-induced stimulation of proliferation by pre-existing cardiomyocytes[29]. Recent evidence indicates that the secreted factor neuregulin1 (Nrg1) is a cardiomyocyte mitogen during cardiac growth or repair in lower and higher vertebrates[30–32]. In zebrafish, *nrg1* is present at very low levels in the heart, and it is induced upon injury at levels that remain undetectable by standard ISH methodology[31]. Strong transgenic overexpression of *nrg1* in adult zebrafish cardiomyocytes activates overt cardiomyocyte proliferation and enlarges the ventricular wall[31]. To test whether *LEN* can influence heart regeneration, we created stable transgenic zebrafish lines with *P2:nrg1* or *LENP2:nrg1* constructs. Resection of the ventricular apex sharply increased *nrg1* transcripts in injured portions of *LENP2:nrg1*, but not control *P2:nrg1*, ventricles (Fig. 6a, b). *LEN*-induced *nrg1* expression was strongest in 7 dpa injury sites, slightly less prominent at 14 dpa, and scarcely detectable by 30 dpa, typically when a contiguous muscle wall has regenerated (Fig. 6a). To examine effects of targeted *nrg1* enhancement, we quantified cardiomyocyte proliferation indices in *LENP2:nrg1* and *P2:nrg1* ventricles at 14 dpa. *LENP2:nrg1* injury sites had a 52% increase in cardiomyocyte proliferation compared to *P2:nrg1* wounds, indicative of improved muscle regeneration (Fig. 6c, d). By 30 dpa, when *nrg1* levels approached baseline, regenerated ventricular walls appeared grossly normal (Fig. 6a). Thus, *LEN* can be designed to deliver mitogenic factors preferentially to areas of cardiac damage, boosting injury-induced cardiomyocyte proliferation.

Discussion

Here, we used a profiling approach to identify small regulatory elements that direct gene expression in regenerating tissue, which we have termed tissue regeneration enhancer elements (TREEs). Recently, a ~18 kb region of the murine *Bmp5* locus was reported to activate expression from minimal promoters in injury contexts[33], suggesting it may harbour a TREE analogous to the *LEN* element we describe here. We suspect that diverse classes of TREEs exist, including elements activated during development and re-activated by injury[34] or during regeneration, elements that activate expression preferentially during regeneration in multiple tissues, and regeneration-specific elements that are more tissue-restricted. The investigation of individual binding motifs within TREEs should identify upstream transcriptional regulators of regeneration, whereas genomic TREE locations can pinpoint novel downstream target genes.

Current methodologies to interrogate regenerative biology often have experimental disadvantages like multiple transgenes, ubiquitous promoters, irreversible expression, and/or stressful stimuli like oestrogen analogues, tetracycline analogues or heat shock[35]. By contrast, TREEs are single-transgene systems that can naturally induce and maintain target genes upon injury, and then naturally temper expression as regeneration concludes. Whereas *LEN* elements induce expression in fin mesenchyme and/or endocardium, we expect that future investigations will uncover a panel of regeneration-responsive TREEs representing additional distinct tissues. Thus, when combined

with effectors or genome-editing enzymes, TREEs should facilitate targeted genetic manipulations that have been elusive to this point.

Multiple features of TREEs are appealing with respect to the design of potential regenerative therapies. Previous studies have implicated the manipulation of enhancer activity as a means to treat human genetic disease[12,36]. In this study, we report that pro- or anti-regenerative factors directed by TREEs are capable of blocking regenerative growth, promoting cell proliferation, or even rescuing genetic defects in regeneration. With a TREE-based system, factor delivery is spatiotemporally defined and could permit therapeutic cycles as injury recurs. Notably, although Nrg1 impacts heart regeneration, systemic neuregulin delivery has the potential for neurological or oncogenic effects[37,38]. Thus, enhancer-based targeting of Nrg1 to injury sites, as we model here in zebrafish, may represent a more effective regenerative medicine platform. We suggest that TREEs identified from natural regenerative contexts across vertebrate species can inform new strategies for precise factor delivery to injured human tissues.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 4 May 2015; accepted 8 March 2016.
Published online 6 April 2016.

1. Poss, K. D. Advances in understanding tissue regenerative capacity and mechanisms in animals. *Nature Rev. Genet.* **11,** 710–722 (2010).
2. Nacu, E. & Tanaka, E. M. Limb regeneration: a new development? *Annu. Rev. Cell Dev. Biol.* **27,** 409–440 (2011).
3. Kumar, A., Godwin, J. W., Gates, P. B., Garza-Garcia, A. A. & Brockes, J. P. Molecular basis for the nerve dependence of limb regeneration in an adult vertebrate. *Science* **318,** 772–777 (2007).
4. Whitehead, G. G., Makino, S., Lien, C. L. & Keating, M. T. *fgf20* is essential for initiating zebrafish fin regeneration. *Science* **310,** 1957–1960 (2005).
5. Wehner, D. & Weidinger, G. Signaling networks organizing regenerative growth of the zebrafish fin. *Trends Genet.* **31,** 336–343 (2015).
6. The ENCODE Project Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
7. Roadmap Epigenomics Consortium Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).
8. Lagha, M., Bothma, J. P. & Levine, M. Mechanisms of transcriptional precision in animal development. *Trends Genet.* **28,** 409–416 (2012).
9. Visel, A., Rubin, E. M. & Pennacchio, L. A. Genomic views of distant-acting enhancers. *Nature* **461,** 199–205 (2009).
10. Giorgio, E. *et al.* A large genomic deletion leads to enhancer adoption by the lamin B1 gene: a second path to autosomal dominant adult-onset demyelinating leukodystrophy (ADLD). *Hum. Mol. Genet.* **24,** 3143–3154 (2015).
11. Rebeiz, M., Pool, J. E., Kassner, V. A., Aquadro, C. F. & Carroll, S. B. Stepwise modification of a modular enhancer underlies adaptation in a *Drosophila* population. *Science* **326,** 1663–1667 (2009).
12. van den Heuvel, A., Stadhouders, R., Andrieu-Soler, C., Grosveld, F. & Soler, E. Long-range gene regulation and novel therapeutic applications. *Blood* **125,** 1521–1525 (2015).
13. Indjeian, V. B. *et al.* Evolving new skeletal traits by *cis*-regulatory changes in bone morphogenetic proteins. *Cell* **164,** 45–56 (2016).
14. Lonfat, N., Montavon, T., Darbellay, F., Gitto, S. & Duboule, D. Convergent evolution of complex regulatory landscapes and pleiotropy at Hox loci. *Science* **346,** 1004–1006 (2014).
15. Herranz, D. *et al.* A NOTCH1-driven *MYC* enhancer promotes T cell development, transformation and acute lymphoblastic leukemia. *Nature Med.* **20,** 1130–1137 (2014).
16. Zhang, Y. *et al.* Positional cloning of the mouse *obese* gene and its human homologue. *Nature* **372,** 425–432 (1994).
17. Fang, Y. *et al.* Translational profiling of cardiomyocytes identifies an early Jak1/Stat3 injury response required for zebrafish heart regeneration. *Proc. Natl Acad. Sci. USA* **110,** 13416–13421 (2013).
18. Kikuchi, K. *et al.* Retinoic acid production by endocardium and epicardium is an injury response essential for zebrafish heart regeneration. *Dev. Cell* **20,** 397–404 (2011).
19. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459,** 108–112 (2009).
20. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl Acad. Sci. USA* **107,** 21931–21936 (2010).
21. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160,** 554–566 (2015).
22. Blow, M. J. *et al.* ChIP-seq identification of weakly conserved heart enhancers. *Nature Genet.* **42,** 806–810 (2010).
23. Simkin, J., Han, M., Yu, L., Yan, M. & Muneoka, K. The mouse digit tip: from wound healing to regeneration. *Methods Mol. Biol.* **1037,** 419–435 (2013).
24. Porrello, E. R. *et al.* Transient regenerative potential of the neonatal mouse heart. *Science* **331,** 1078–1080 (2011).
25. Park, C., Kim, T. M. & Malik, A. B. Transcriptional regulation of endothelial cell and vascular development. *Circ. Res.* **112,** 1380–1400 (2013).
26. De Val, S. *et al.* Combinatorial regulation of endothelial gene expression by ets and forkhead transcription factors. *Cell* **135,** 1053–1064 (2008).
27. Lee, Y., Grill, S., Sanchez, A., Murphy-Ryan, M. & Poss, K. D. Fgf signaling instructs position-dependent growth rate during zebrafish fin regeneration. *Development* **132,** 5173–5183 (2005).
28. Amaya, E., Musci, T. J. & Kirschner, M. W. Expression of a dominant negative mutant of the FGF receptor disrupts mesoderm formation in *Xenopus* embryos. *Cell* **66,** 257–270 (1991).
29. Kikuchi, K. *et al.* Primary contribution to zebrafish heart regeneration by *gata4*$^{+}$ cardiomyocytes. *Nature* **464,** 601–605 (2010).
30. Polizzotti, B. D. *et al.* Neuregulin stimulation of cardiomyocyte regeneration in mice and human myocardium reveals a therapeutic window. *Sci. Translat. Med.* **7,** 281ra245 (2015).
31. Gemberling, M., Karra, R., Dickson, A. L. & Poss, K. D. Nrg1 is an injury-induced cardiomyocyte mitogen for the endogenous heart regeneration program in zebrafish. *eLife* **4,** e05871 (2015).
32. Bersell, K., Arab, S., Haring, B. & Kuhn, B. Neuregulin1/ErbB4 signaling induces cardiomyocyte proliferation and repair of heart injury. *Cell* **138,** 257–270 (2009).
33. Guenther, C. A. *et al.* A distinct regulatory region of the *Bmp5* locus activates gene expression following adult bone fracture or soft tissue injury. *Bone* **77,** 31–41 (2015).
34. Huang, G. N. *et al.* C/EBP transcription factors mediate epicardial activation during heart development and injury. *Science* **338,** 1599–1603 (2012).
35. Gemberling, M., Bailey, T. J., Hyde, D. R. & Poss, K. D. The zebrafish as a model for complex tissue regeneration. *Trends Genet.* **29,** 611–620 (2013).
36. Deng, W. *et al.* Reactivation of developmentally silenced globin genes by forced chromatin looping. *Cell* **158,** 849–860 (2014).
37. Nawa, H., Sotoyama, H., Iwakura, Y., Takei, N. & Namba, H. Neuropathologic implication of peripheral neuregulin-1 and EGF signals in dopaminergic dysfunction and behavioral deficits relevant to schizophrenia: their target cells and time window. *Biomed. Res. Int.* **2014,** 697935 (2014).
38. Montero, J. C. *et al.* Neuregulins and cancer. *Clin. Cancer Res.* **14,** 3237–3241 (2008).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. Burris, N. Lee, and T. Thoren for zebrafish care; A. Knecht, and J. Savage for technical advice or assistance; and M. Bagnat, C. Chen, F. Conlon, D. Fox, and M. Mokalled for comments on the manuscript. J.H. was supported by an AHA postdoctoral fellowship (12POST11920060), R.K. by an NIH Clinical Investigator Award (K08 HL116485), V.A.T. by an NSF Graduate Research Fellowship (1106401), and J.A.G. by an NIH postdoctoral fellowship (F32 HL120494). This work was supported by NIH grants to B.L.B. (R01 HL089707 and R01 HL064658) and K.D.P. (R01 GM074057 and R01 HL081674), who acknowledges support from HHMI.

Author Contributions J.K. and K.D.P. designed the experimental strategy, analysed data, and prepared the manuscript. J.K. generated transgenic zebrafish and performed regeneration experiments and analysis. J.H. and B.L.B. generated and analysed transgenic mice. R.K. generated and analysed sequencing datasets. A.L.D. performed surgeries and histology. V.A.T. generated and analysed mutant zebrafish. G.N., M.G. and J.A.G. contributed unpublished results and methodology. All authors commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to K.D.P. (kenneth.poss@duke.edu).

METHODS

Zebrafish maintenance and procedures. Wild-type or transgenic male and female zebrafish of the outbred Ekkwill (EK) strain were used for all experiments, with adults ranging in age from 3 to 12 months. Water temperature was maintained at 26°C for animals unless otherwise indicated. Fins were amputated to 50% of their original length using razor blades. As penetrance of the *dob* mutation was higher at 33°C than at 26°C, *dob* fish were maintained at 33°C after caudal fin amputation. To measure lengths of regenerates, lengths from the amputation plane to the distal tips of the third and fourth fin rays of dorsal and ventral caudal fin lobes were determined using ZEN software. Because some *dob* animals regenerated portions of the first and second fin rays of ventral lobes, regenerating caudal fin areas for Extended Data Fig. 10c were measured from the dorsal third fin ray to the ventral third fin ray and calculated using ZEN software. Partial ventricular resection surgeries were performed as described previously[39], in which ~20% of the cardiac ventricle was removed at the apex. To ablate cardiomyocytes, *cmlc2:CreER; bactin2:loxp-mCherry-STOP-loxp-DTA* (Z-CAT) fish were used[40]. Z-CAT zebrafish were incubated in vehicle (0.01% EtOH) or 10 μM tamoxifen for 12 h. Work with zebrafish was performed in accordance with Duke University guidelines.

To generate *lepb:eGFP* BAC transgenic animals (full names, *Tg(lepb:eGFP)^{pd120} and Tg(lepb:eGFP)^{pd121}*), the *iTol2* cassette[41] was integrated into the BAC clone DKEY-21O22 using Red/ET recombineering technology (GeneBridges). Then, the first exon of the *lepb* gene in the BAC clone DKEY-21O22 was replaced with an *eGFP* cassette by Red/ET recombineering. 5′ and 3′ homology arms were amplified by PCR (Supplementary Information) and subcloned into the pCS2-eGFP plasmid. One nl of 50 ng μl^{-1} purified, recombined BAC was injected into one-cell stage zebrafish embryos along with one nl of 30 ng μl^{-1} synthetic *Tol2* mRNA[41]. To sort F_0 transgenic animals injected with *lepb:eGFP* constructs, fin folds were amputated at 3 or 4 dpf, and embryos displaying eGFP fluorescence near the injury site at 1 dpa were selected (Extended Data Fig. 1f). After raising F_0 zebrafish to adulthood, caudal fins were amputated and zebrafish displaying induced eGFP were selected for breeding (Extended Data Fig. 1g). Between 30–60 dpf, caudal fins of progeny from transgene-positive F_0 fish were amputated, and eGFP$^+$ transgenic animals were isolated to identify stable lines. Two lines were identified that had indistinguishable expression features.

To define *LEN* activity, over 60 additional new transgenic lines were established in this study, listed in Supplementary Data 1. To generate transgenic animals, DNA sequences were amplified by PCR with indicated primers (Supplementary Data 3) and subcloned into a pCS2-eGFP-I-sceI vector, in which I-SceI restriction sites were flanked by a multiple cloning site. As promoters, 2 kb, 1.6 kb, and 0.7 kb upstream sequences of *lepb*, *cmlc2* (ref. 42), and α-*cry*[43] genes were used, respectively. These constructs were injected into one-cell-stage wild-type or *dob* embryos using standard meganuclease transgenesis techniques. 2 kb *lepb* upstream sequences could induce transgene expression after fin fold amputation at larval stages, but never after caudal fin amputation in adults. To isolate stable lines, larvae were examined for transgene expression near injury site in response to fin fold amputation (2 kb *lepb*), in cardiomyocytes (1.6 kb *cmlc2*), and in lens (0.7 kb α-*cry*).

To test additional TREEs, we subcloned putative enhancer regions of *il11a, plek, vcana,* and *cd248b* upstream of 800 bp of *lepb* upstream sequence (P0.8). To define TREE activity, these constructs were injected into one-cell-stage wild-type embryos, Fin folds were amputated at 4 dpf, and eGFP fluorescence near the amputation plane was examined at 5 dpf (1 dpa).

Generation and analysis of transgenic mice. Transgenic mice (CD-1 strain) were generated by oocyte microinjection as described previously[44]. *LEN-hsp68::lacZ* transgenic mice were generated by subcloning the zebrafish *LEN* enhancer sequence into the transgenic reporter plasmid *hsp68-lacZ*[45]. Ctrl-*hsp68::lacZ* transgenic mice harbour a transgene, *Prkaa2*[mMEF2(1+2)]*-hsp68-lacZ*, which contains a modified version of a 931-bp enhancer sequence from the mouse *Prkaa2* gene cloned into *hsp68-lacZ* (J. Hu and B. L. Black, unpublished observations). Partial apical resection injury in male and female neonatal mice at postnatal day 1 was performed similarly to previously described methods[46]. Hearts and paws were collected at postnatal day 4. All experiments with mice complied with federal and institutional guidelines and were reviewed and approved by the UCSF IACUC.

RNA isolation and quantitative PCR. RNA was isolated from dissected caudal fins and partially resected ventricles using Tri-Reagent (Sigma). cDNA was synthesized from 1 μg of total RNA using the Roche First Strand Synthesis Kit. Quantitative PCR was performed using the Roche LightCycler 480 and the Roche LightCycler 480 Probes Master. All samples were analysed in biological triplicates and technical duplicates. Primer sequences are described in Supplementary Information, and probe numbers for *actb2, lepb,* and *nrg1* were 104, 156 and 76, respectively. *lepb* and *nrg1* transcript levels were normalized to *actb2* levels for all experiments.

RNA sequencing. Total RNA was prepared from two biological replicate pools of ablated Z-CAT ventricles and uninjured ventricles at 7 days post-ablation as per Gemberling *et al.*[31], or regenerating and uninjured caudal fins. Generation of mRNA libraries and sequencing were performed at the Duke Genome Sequencing Shared Resource using an Illumina HiSeq2000. Sequences were aligned to the zebrafish genome (Zv9) using TopHat[47]. Differentially regulated transcripts were identified using EdgeR and an FDR cut-off of 0.1 (ref. 48). Accession numbers for transcriptome data sets are GSE75894 and GSE76564.

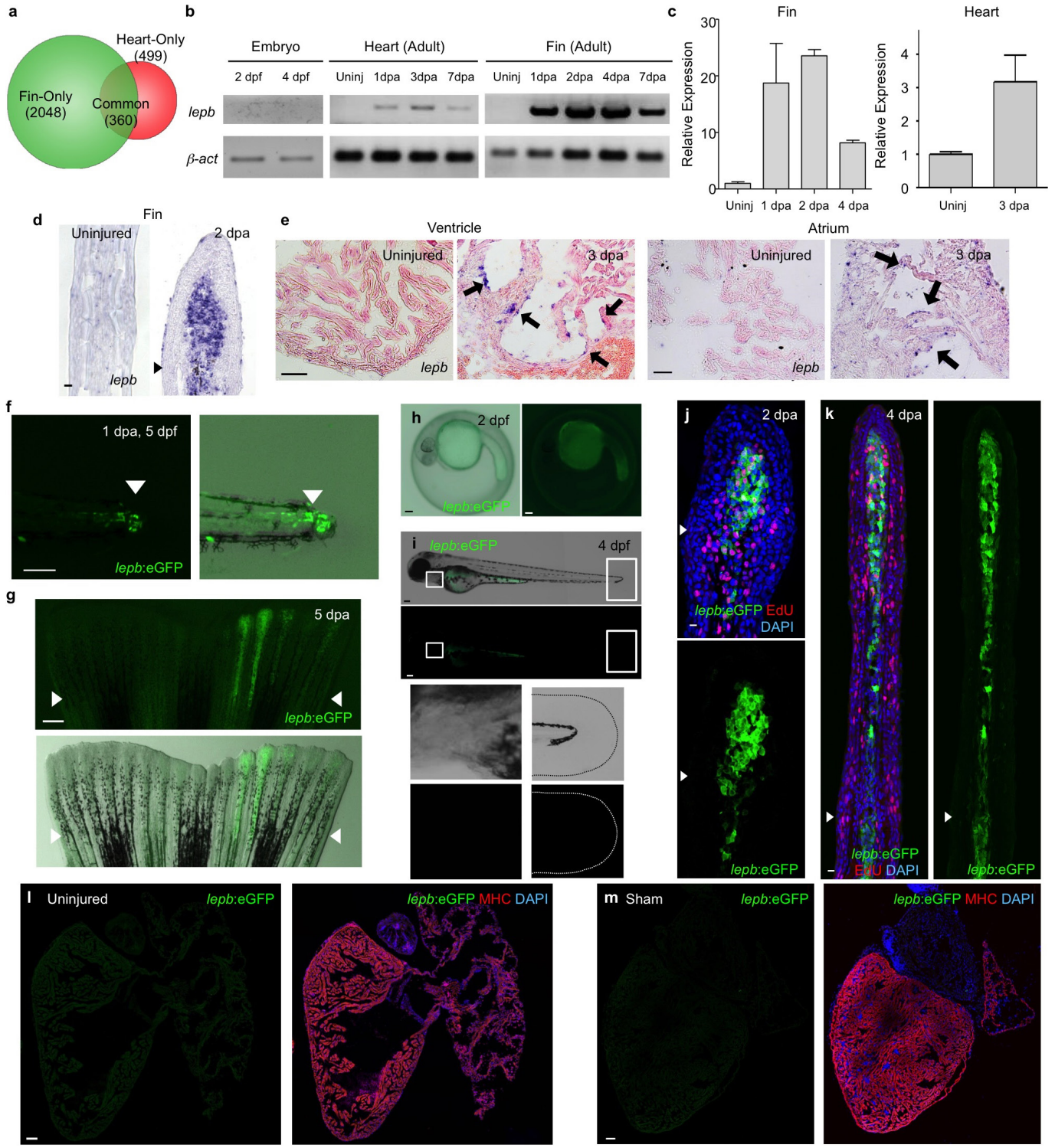
ChIP sequencing. To identify candidate enhancer elements activated during heart regeneration, chromatin extracts were prepared from two biological replicate pools of 10 ablated Z-CAT ventricles and 10 uninjured ventricles. Chromatin was sonicated and immunoprecipitated with an antibody against H3K27ac (ActiveMotif) using the MAGnify ChIP system (Invitrogen). Sequencing libraries were prepared as per Bowman, *et al.*[49]. Sequencing was performed using an Illumina HiSeq2000, and 10–25 million 50 bp single end reads were obtained for each library. Sequences were aligned to the zebrafish genome (Zv9) using Bowtie2 (ref. 50). Differential peaks were identified using Model-based Analysis for ChiP-seq (MACS)[51].

Histology and imaging. *In situ* hybridization on cryosections of 4% paraformaldehyde-fixed fins was performed as described previously[52]. To generate digoxigenin-labelled probes for *lepb* and *fgf20a*, we generated a fragment of *lepb* cDNA and a full length of *fgf20a* cDNA by PCR using primer sequences described in Supplementary Information. The *nrg1* probe was prepared as described previously[31]. Immunohistochemistry was performed as described previously[40]. Primary and secondary antibodies used in this study were: anti-myosin heavy chain (mouse, F59, Developmental Studies Hybridoma Bank), anti-MEF2 (rabbit, sc-313, Santa Cruz Biotechnology), anti-PCNA (mouse, P8825, Sigma), anti-eGFP (rabbit, A11122, Life Technologies), anti-eGFP (chicken, GFP-1020, Aves Labs), anti-Raldh2 (rabbit, Abmart), anti-Ds-Red (rabbit, 632496, Clontech), anti-p63 (mouse, 4A4, Santa Cruz Biotechnology), Alexa Fluor 488 (mouse and rabbit; Life Technologies), Alexa Fluor 594 (mouse and rabbit; Life Technologies). For EdU incorporation experiments, zebrafish were injected intraperitoneally with 10 mM EdU (A10055, sigma), and caudal fins were collected at 1 h post-treatment. EdU staining was performed as previously described[53]. The secondary antibody used for EdU staining was Alexa 488 azide (10–20 μM, Sigma). Whole-mount images were acquired using an M205FA stereofluorescence microscope (Leica) or Axio Zoom (Zeiss). Images of tissue sections (10 μm for hearts and 14 μm for fins) were acquired using an LSM 700 confocal microscope (Zeiss). X-gal staining to detect β-galactosidase activity and counterstaining with nuclear fast red were performed with murine tissue as described previously[44].

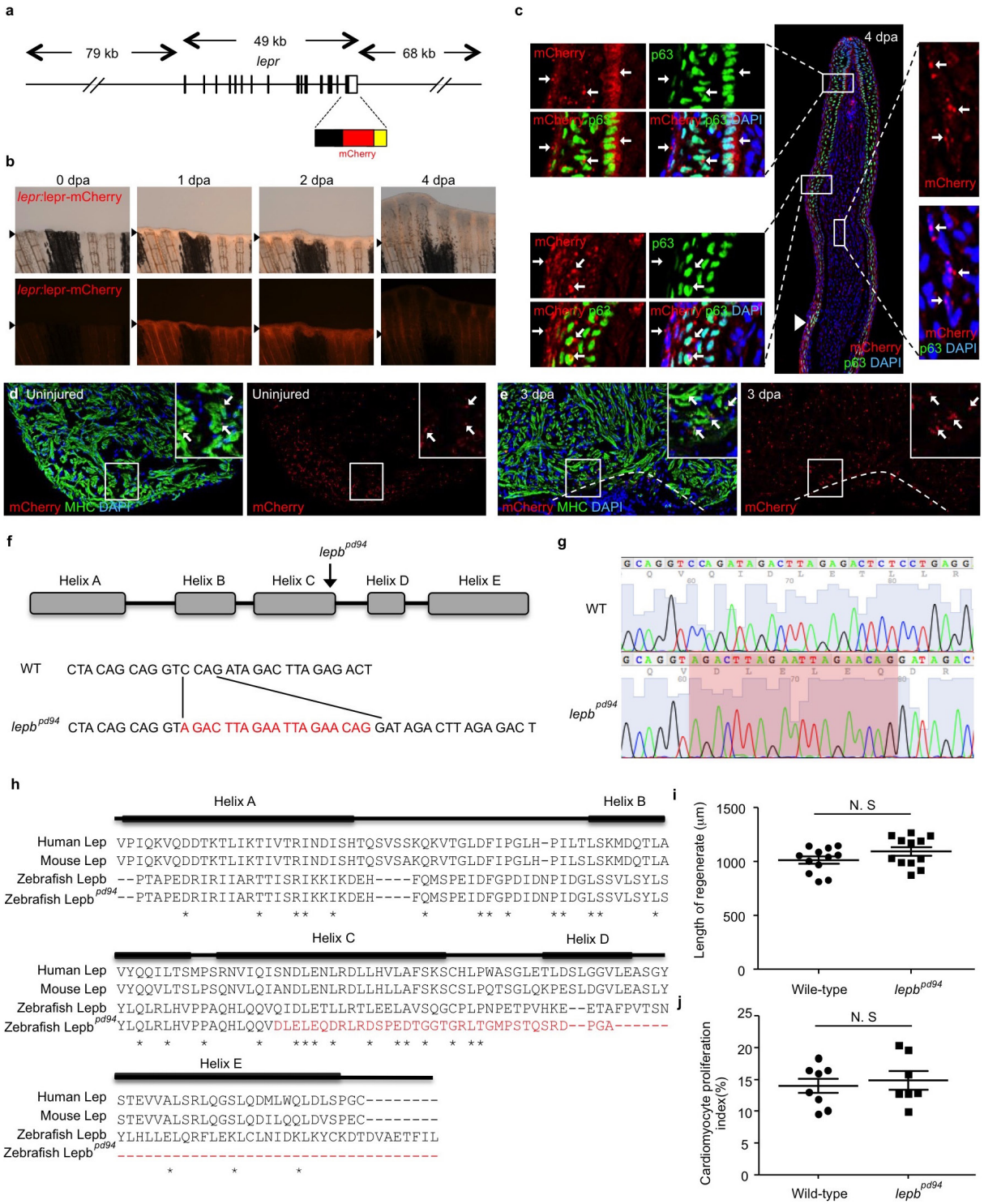
Data collection and statistics. Clutchmates were randomized into different treatment groups for each experiment. No animal or sample was excluded from the analysis unless the animal died during the procedure. Sample sizes were chosen on the basis of previous publications and experiment types, and are indicated in each figure legend or methods. No statistical methods were used to predetermine sample size. For expression patterns, at least five fish from each transgenic line were examined. At least 9 hearts of each group were pooled for RNA purification and subsequent RT–qPCR. Quantification of cardiomyocyte proliferation and calculation of statistical outcomes were assessed by a person blinded to the treatments. Other experiments were not blinded during experiments and outcome assessment. Sample sizes, statistical tests, and *P* values are indicated in the figures or the legends. One-way ANOVA tests were applied when normality and equal variance tests were passed. The Mann–Whitney rank sum test was applied in assays of cardiomyocyte proliferation.

39. Poss, K. D., Wilson, L. G. & Keating, M. T. Heart regeneration in zebrafish. *Science* **298,** 2188–2190 (2002).
40. Wang, J. *et al.* The regenerative capacity of zebrafish reverses cardiac failure caused by genetic cardiomyocyte depletion. *Development* **138,** 3421–3430 (2011).
41. Suster, M. L., Abe, G., Schouw, A. & Kawakami, K. Transposon-mediated BAC transgenesis in zebrafish. *Nature Protocols* **6,** 1998–2021 (2011).
42. Burns, C. G. *et al.* High-throughput assay for small molecules that modulate zebrafish embryonic heart rate. *Nature Chem. Biol.* **1,** 263–264 (2005).
43. Kurita, R. *et al.* Suppression of lens growth by αA-crystallin promoter-driven expression of diphtheria toxin results in disruption of retinal cell organization in zebrafish. *Dev. Biol.* **255,** 113–127 (2003).
44. Dodou, E., Xu, S. M. & Black, B. L. *mef2c* is activated directly by myogenic basic helix-loop-helix proteins during skeletal muscle development in vivo. *Mech. Dev.* **120,** 1021–1032 (2003).
45. Kothary, R. *et al.* Inducible expression of an hsp68–lacZ hybrid gene in transgenic mice. *Development* **105,** 707–714 (1989).

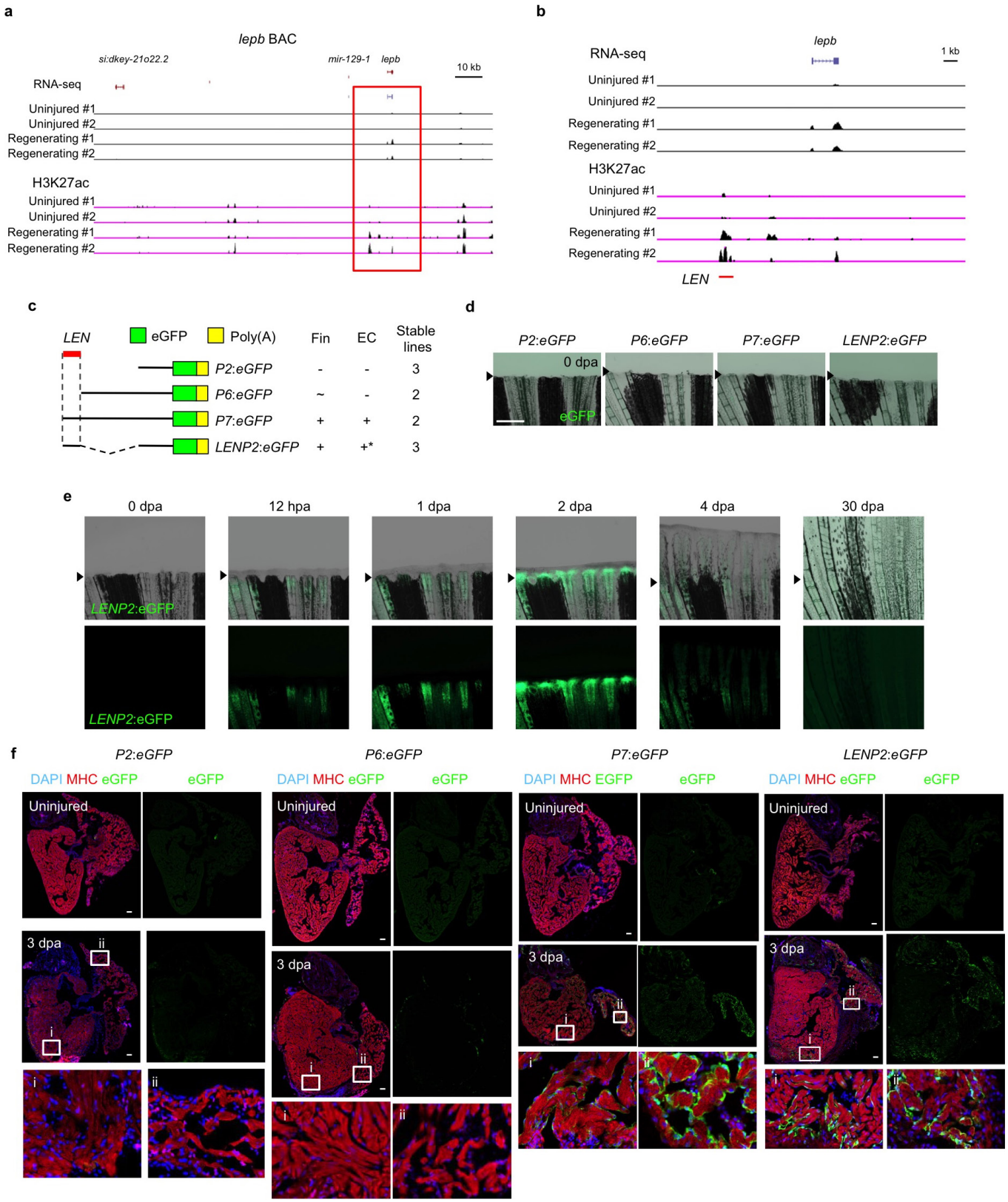
46. Mahmoud, A. I., Porrello, E. R., Kimura, W., Olson, E. N. & Sadek, H. A. Surgical models for cardiac regeneration in neonatal mice. *Nature Protocols* **9,** 305–311 (2014).
47. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).
48. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).
49. Bowman, S. K. *et al.* Multiplexed Illumina sequencing libraries from picogram quantities of DNA. *BMC Genomics* **14,** 466 (2013).
50. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9,** 357–359 (2012).
51. Zhang, Y. *et al.* Model-based analysis of ChIP-seq (MACS). *Genome Biol.* **9,** R137 (2008).
52. Lee, Y. *et al.* Maintenance of blastemal proliferation by functionally diverse epidermis in regenerating zebrafish fins. *Dev. Biol.* **331,** 270–280 (2009).
53. Salic, A. & Mitchison, T. J. A chemical method for fast and sensitive detection of DNA synthesis *in vivo. Proc. Natl Acad. Sci. USA* **105,** 2415–2420 (2008).



Extended Data Figure 1 | *lepb* transcripts sharply increase during fin and heart regeneration. a, Venn diagram displaying numbers of genes with significantly increased transcript levels during fin and heart regeneration. **b**, RT–PCR of samples from 2 days post-fertilization (dpf) and 4 dpf embryos, and uninjured and regenerating adult tissues. *lepb* was not detected during embryogenesis and in uninjured tissues, but induced during regeneration. *β-act2* is used as loading control. Uninj, Uninjured. **c**, Left: relative expression of *lepb* in uninjured, 1, 2, and 4 dpa fin regenerates. *lepb* transcript levels are increased at 1 and 2 dpa. Right: relative expression of *lepb* in uninjured or 3 dpa cardiac ventricles, assessed by qPCR. **d**, **e**, Endogenous *lepb* expression assessed by *in situ* hybridization in sections of fins (**d**) and cardiac ventricle and atrium (**e**). Arrowhead, amputation plane. Arrows, endocardial *lepb* expression. Left: uninjured tissues, Right: regenerating tissues. dpa: days post-amputation. **f**, **g**, F_0 animals, injected with the transgenic *lepb*:*eGFP* BAC reporter construct at the one-cell stage, induced eGFP after larval fin fold amputation (**f**) and during adult fin regeneration (**g**). Note that *lepb*:eGFP is mosaically expressed. Arrowheads, amputation planes. **h**, **i,** Expression pattern of *lepb*:*eGFP* stable transgenic animals. *lepb*:eGFP was not detected in fin and heart during embryogenesis (2 dpf (**h**); 4 dpf (**i**)). Below '**i**' are enlargements of the boxed areas, which show heart (left) and fin fold (right). Dotted line, outline of fin fold. The yolk is autofluorescent. **j**, **k**, Section images of *lepb*:*eGFP* caudal fin regenerates at 2 dpa (**j**) and 4 dpa (**k**). The majority of *lepb*:eGFP-positive cells are mesenchymal cells, overlapping partially with cells that incorporate EdU (collected 60 min after injection; red). **l**, **m**, Lack of detectable expression of *lepb*:eGFP in hearts of uninjured (**l**) or sham-operated (**m**) *lepb*:*eGFP* animals. $n = 8$ and 5 for uninjured and sham-operated hearts, respectively. Arrowheads, amputation planes. Scale bars represent 10 μm (**d**, **f**, **h**–**k**); 50 μm (**e**, **l**, **m**); 500 μm (**g**).

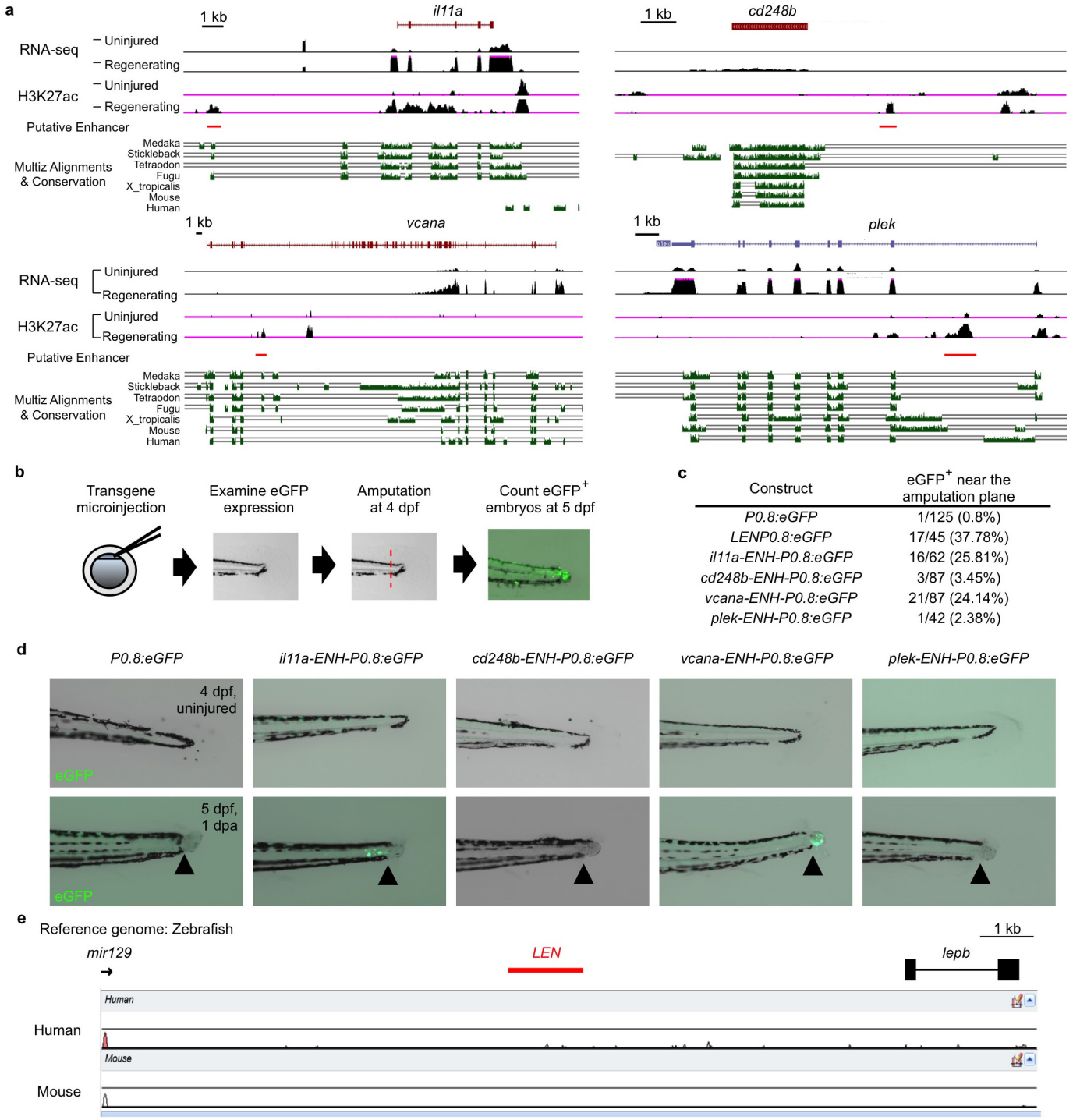


Extended Data Figure 2 | Leptin signalling during fin and heart regeneration. a–**e**, Expression pattern of *lepr:lepr–mCherry* BAC reporter line. **a**, Schematic of the *lepr:lepr–mCherry* BAC transgenic construct. mCherry is fused at the C terminus of Lepr. **b**, mCherry fluorescence in the *lepr:lepr–mCherry* BAC reporter strain is induced during fin regeneration. $n = 5$; all animals displayed a similar expression pattern. **c**, Section images of 4 dpa *lepr:lepr–mCherry* caudal fin regenerates. The majority of Lepr–mCherry^{+} cells are epidermal cells, overlapping partially with p63^{+} basal and suprabasal cells (left). In addition, some putative vascular cells in the intra-ray region have Lepr–mCherry signals (right). **d**, **e**, Confocal images of sections through uninjured (**d**) and regenerating (**e**) *lepr:lepr–mCherry* hearts. Lepr–mCherry fluorescence co-localizes with MHC^{+} cardiomyocytes in uninjured and 3 dpa hearts (arrows). Note that these expression patterns are similar to leptin receptor expression in mice (see Supplementary Information). $n = 7$ and 6 for uninjured and 3 dpa hearts, respectively. **f**–**j**, Analysis of fin and heart regeneration in *lepb*^{pd94} mutants. **f**, A schematic representation of Lepb, showing the effects of the *pd94* mutation. Lepb is composed of 5 alpha-helix domains. *lepb*^{pd94} has a 19 bp insertion and a 3 bp deletion at the third α-helix (helix C). **g**, Sequencing of wild-type and *lepb*^{pd94} alleles revealed an indel (red highlight). **h**, A comparison of the amino acid sequences of leptin genes of human, mice, and zebrafish. The predicted amino acid sequence of the *lepb*^{pd94} gene product is shown at the bottom, with the predicted truncation sites indicated in red. The predicted *lepb*^{pd94} protein product lacks the majority of C-terminal amino acids. Asterisk indicates identical amino acid residue between three species. **i**, Quantification of regenerated fin lengths from *lepb*^{pd94} and wild type siblings at 4 dpa. $n = 12$ each of *lepb*^{pd94} and wild-type. **j**, Quantification of cardiomyocyte proliferation at 7 dpa. $n = 7$ (*lepb*^{pd94}) and 8 (wild-type). Data are represented as mean ± s.e.m. NS, not significant.

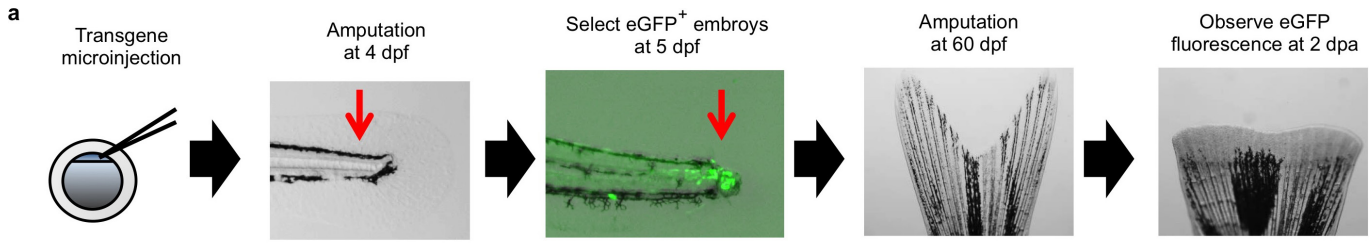


Extended Data Figure 3 | See next page for caption.

Extended Data Figure 3 | Identification of *LEN* and tests of regulatory sequences near *lepb*. **a**, Schematic depicting the genomic region surrounding *lepb* (corresponding to the *lepb* BAC used in this study) with the profiles of RNA-sequencing and H3K27ac marks from uninjured and regenerating heart tissues. **b**, Enlargement of the boxed area in **a**. *lepb* is the only upregulated gene in this genomic region during regeneration. H3K27ac-enriched peaks in regenerating samples are present in a ~1 kb region (red bar) that is ~7 kb upstream of the start codon. **c**, Schematic representation of transgenes to examine regulatory sequence activity. Fin and endocardial expression during regeneration and the number of stable lines are indicated. Asterisk indicates that one *LENP2:eGFP* line showed occasional, weak endocardial eGFP expression in uninjured hearts, whereas eGFP signal in this line was broad and strong during regeneration. EC, endocardial cells. **d**, Images of representative 0 dpa fins from lines indicated in **c**. eGFP fluorescence is not detectable in fins at 0 dpa or in uninjured fins, but is induced in regenerating ray blastemas in *P7:eGFP* and *LENP2:eGFP* lines. *P6:eGFP* regenerates displayed weak eGFP expression below the amputation plane during regeneration, with very weak or undetectable expression in regenerating portions (see Fig. 2c). **e**, *LENP2*:eGFP expression pattern during fin regeneration. eGFP is detectable as early as 12 hpa, but is undetectable at 30 dpa. $n = 5$; all animals displayed a similar expression pattern. Arrowheads, amputation planes. **f**, Section images of representative uninjured and regenerating hearts from *P2:eGFP*, *P6:eGFP*, *P7:eGFP*, and *LENP2:eGFP* animals. eGFP fluorescence is rarely detectable in uninjured *P2:eGFP*, *P6:eGFP*, *P7:eGFP*, or *LENP2:eGFP* hearts, except in one line of *LENP2:eGFP* (mentioned above). Upon injury, *P2* drove weak, occasional expression in cardiomyocytes and epicardium but not in endocardium, whereas *P7* and *LEN* drove endocardial eGFP expression in ventricle and atrium. i, ii, enlargements of boxes areas in regenerating ventricle and atrium, respectively. Scale bars: 500 μm (**d**, **e**); 50 μm (**f**).



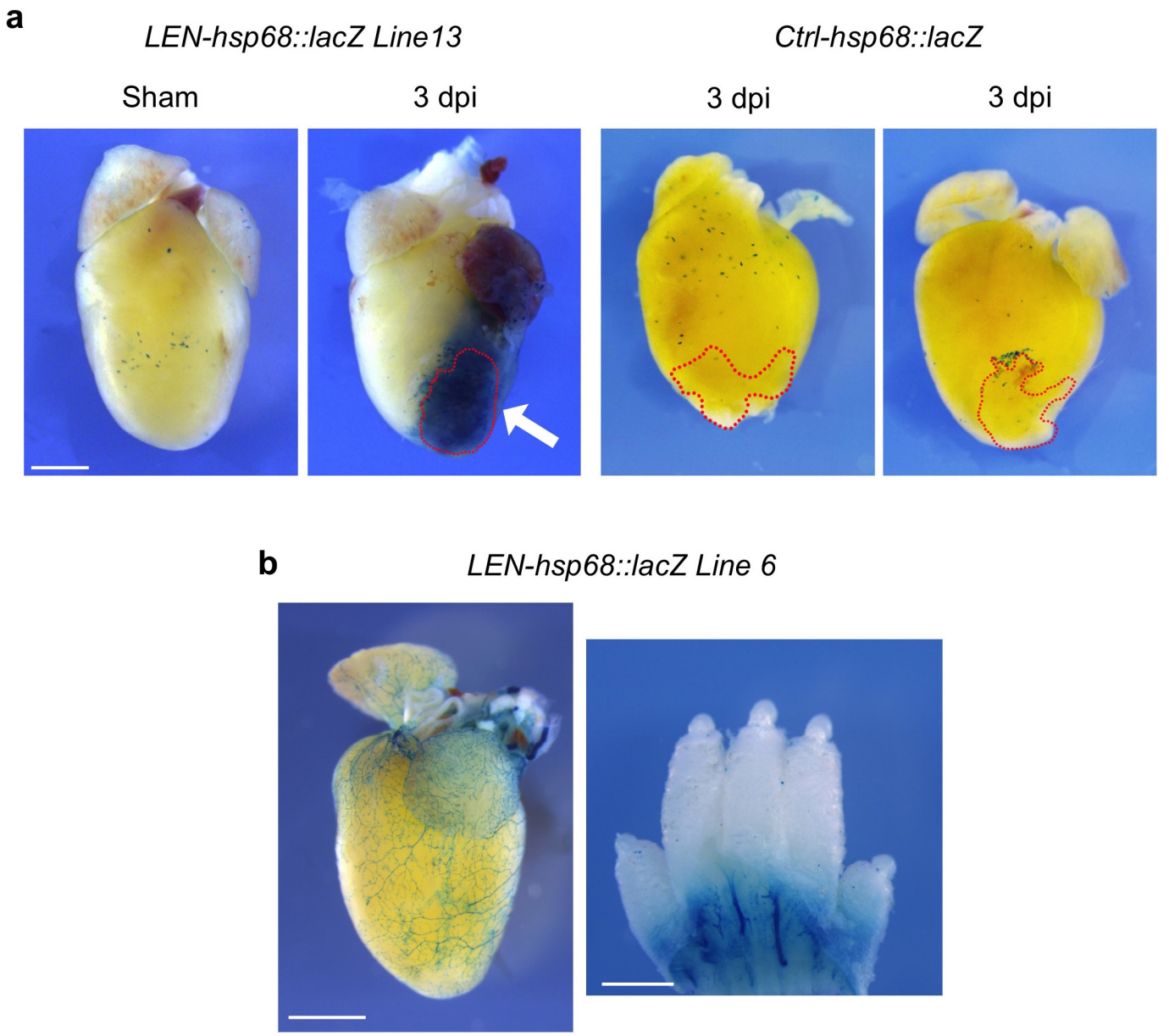
Extended Data Figure 4 | Additional putative regeneration enhancer elements. a, Cartoon depicting the distal upstream regions of *il11a, cd248b, vcana,* and *plek*. RNA-sequencing profiles indicate that these genes are upregulated during heart regeneration. The red bar indicates putative enhancer regions that are enriched with H3K27ac marks in regenerating tissue. Two of these putative enhancers, near *il11a* and *vcana*, showed primary sequence conservation in other non-mammalian vertebrates but not in mammals. **b**, Scheme depicting assays in injected F_0 transgenic animals. At 4 dpf, eGFP expression in the uninjured fin fold was examined, and then the fin fold was amputated. eGFP expression near the amputation plane was examined at 5 dpf. **c**, Table indicating injected constructs and the number of animals with $eGFP^+$ cells near the amputation plane. **d**, Images of representative 4 dpf (uninjured) and 5 dpf (regenerating) fin folds from animals in **c**. **e**, Vista plot of genomic regions from *mir129* to *lepb* based on *LAGAN* alignment with reference sequence zebrafish. Sequence comparison indicates that this region is not highly conserved between zebrafish and mammals. Arrowheads, amputation planes.



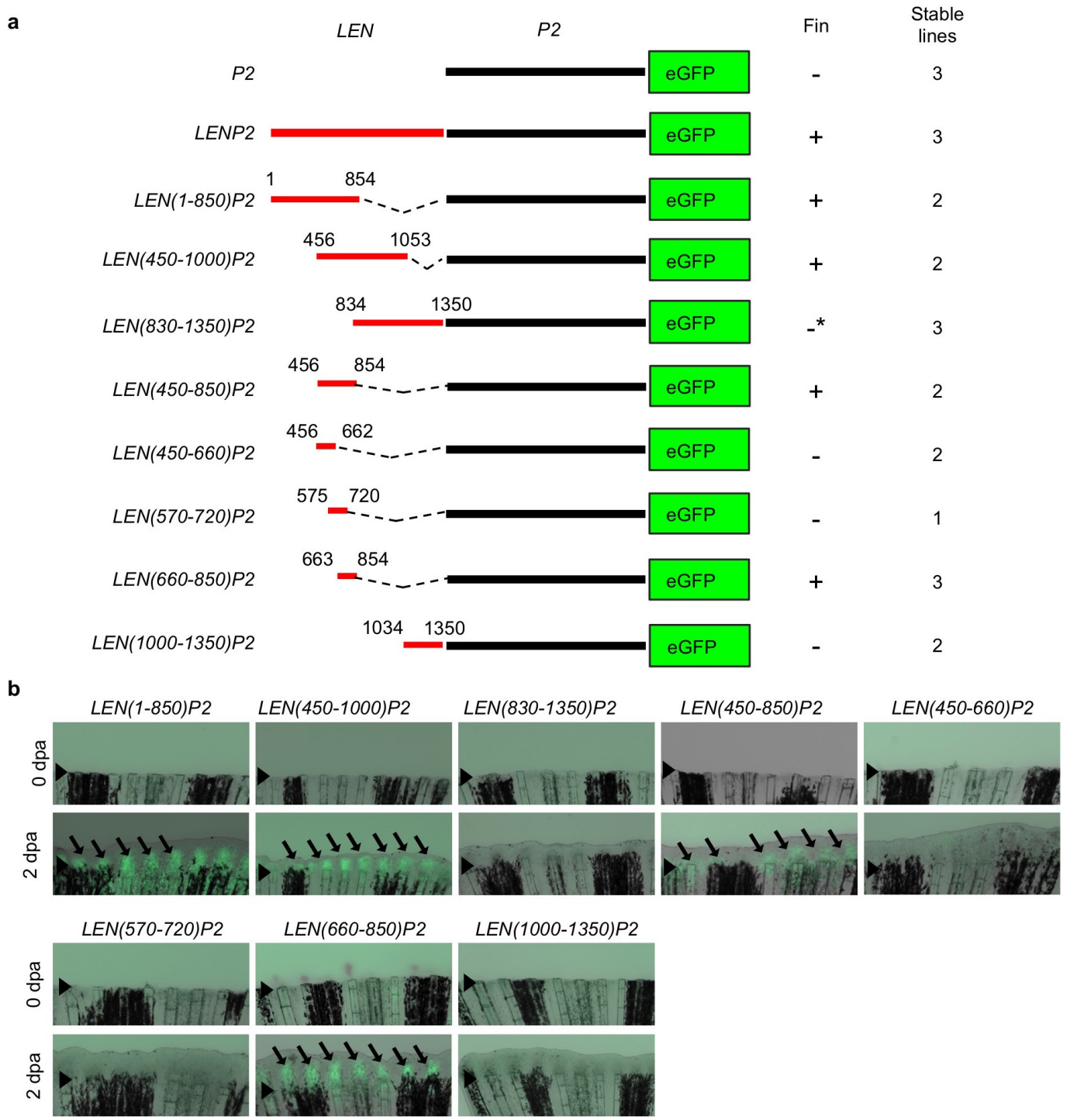
b

	LEN Promoter	lepb P2	cmlc2	α-cry
LENpromoter	eGFP	N.D	8/38	1/11
LEN(1-450)promoter	1 463 eGFP	0/12	N.D	N.D
LEN(1-850)promoter	1 854 eGFP	7/24	6/28	0/5
LEN(450-1000)promoter	456 1053 eGFP	7/30	16/40	0/91
LEN(830-1350)promoter	834 1350 eGFP	1/33	0/24	0/3
LEN(450-850)promoter	456 854 eGFP	4/28	3/31	N.D
LEN(850-1000)promoter	854 1053 eGFP	0/22	N.D	N.D
LEN(1000-1350)promoter	1034 1350 eGFP	0/73	0/29	N.D
LEN(450-660)promoter	456 662 eGFP	0/42	0/9	N.D
LEN(570-720)promoter	575 720 eGFP	0/4	0/14	N.D
LEN(660-850)promoter	663 854 eGFP	4/20	1/5	N.D

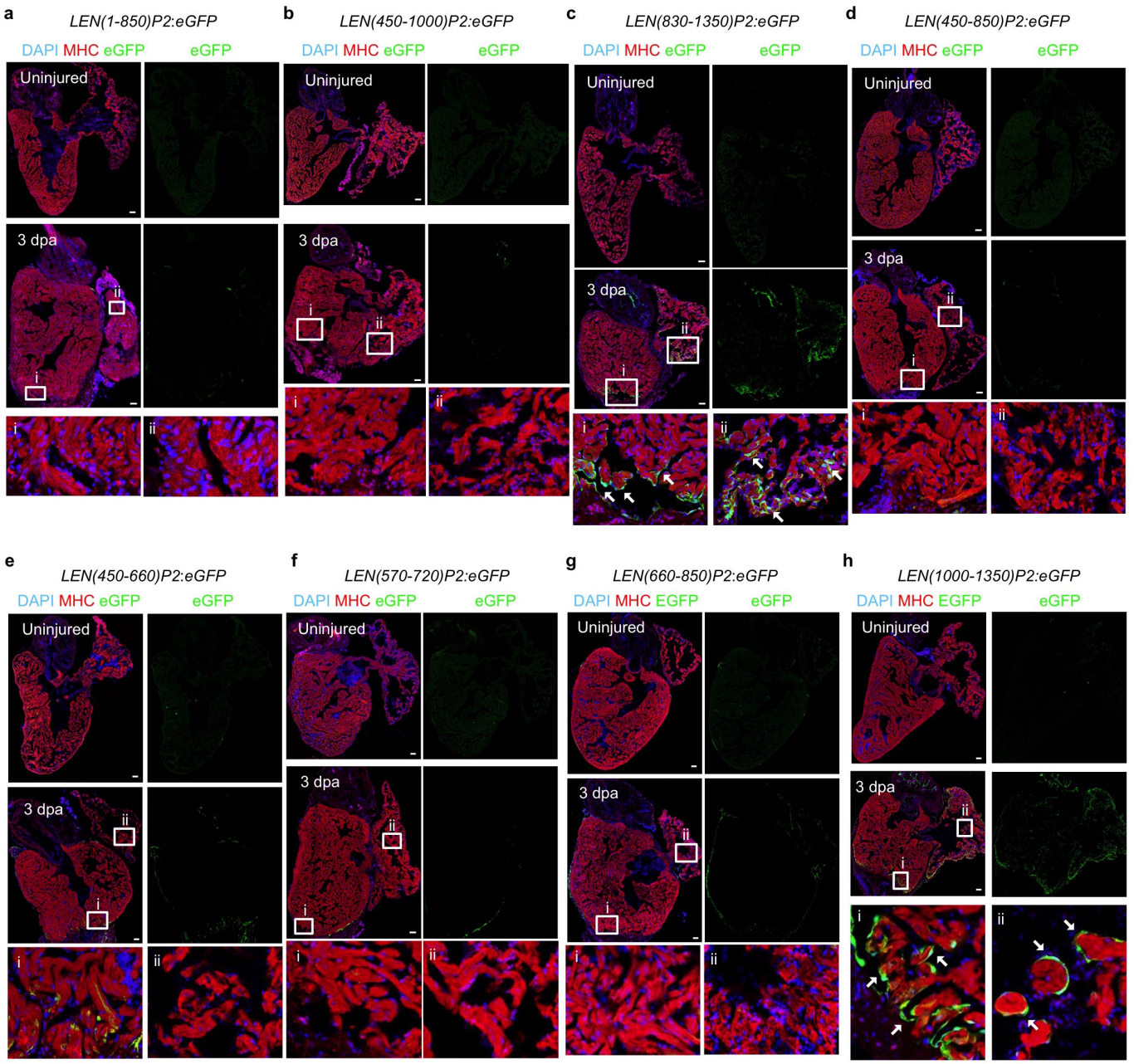
Extended Data Figure 5 | Transient transgenic assays examining *lepb*-linked regeneration enhancer fragments in combination with different promoters (fin regeneration). a, Scheme depicting assays in injected F_0 transgenic animals. Transgene-positive larvae were selected by detection of eGFP in response to fin fold amputation (*lepb* promoter), in cardiomyocytes (*cmlc2* promoter), or in lenses (*α-cry* promoter). Caudal fins of F_0 transgenic positive zebrafish were amputated at 60–90 days post-fertilization (dpf), and eGFP expression was examined at 2 dpa. **b**, Schematic representation of the transgenic constructs to examine fin regeneration enhancer activity. Expression during fin regeneration and the number of assessed F_0 animals are indicated. Many embryos transgenic for *LEN*(1–850), *LEN*(450–1000), *LEN*(450–850), and *LEN*(660–850) coupled with the *lepb* or *cmlc2* promoter showed activity during fin regeneration. One of eleven *LENα-cry:eGFP* animals displayed fin eGFP expression, but *LEN*(1–850)*α-cry:eGFP* and *LEN*(450–1000)*α-cry:eGFP* did not drive eGFP expression during fin regeneration, indicating that there may be repressive motifs in the *α-cry* promoter fragment that affect fin regeneration enhancer activity (See also Extended Data Fig. 9). ND, not determined.



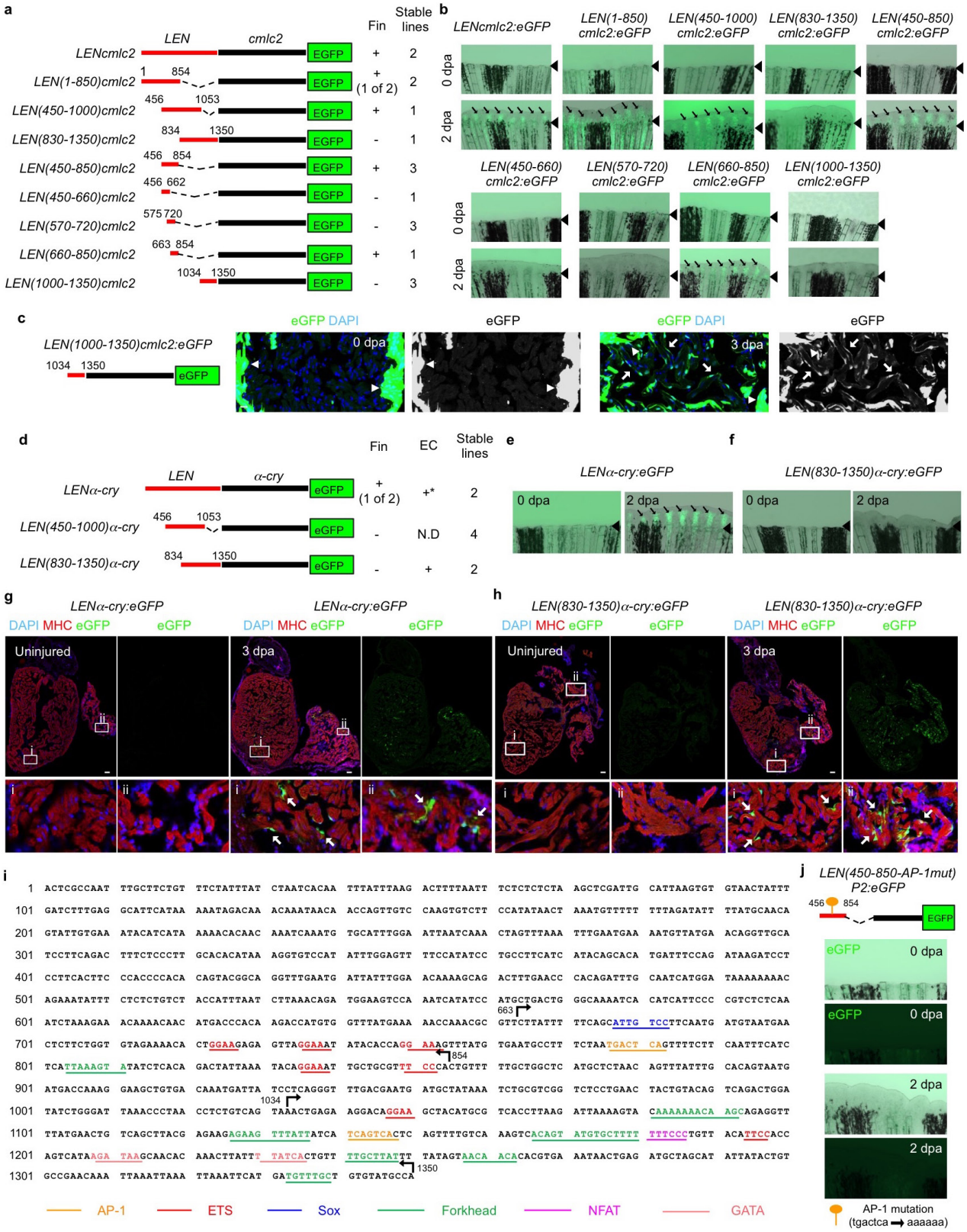
Extended Data Figure 6 | X-gal staining in stable transgenic mouse lines. a, Additional whole mount images of X-gal stained hearts from neonatal *LEN-hsp68::lacZ* (line 13, presented in Fig. 3) and control animals injured at postnatal day 1 and assessed at postnatal day 4. X-gal staining is undetectable in sham-operated hearts of *LEN-hsp68::lacZ* mice ($n = 6$; representative image shown) and injured hearts of control mice, but strong in partially resected hearts of *LEN-hsp68::lacZ* mice (arrows). Dashed red lines indicate injury area, positioned facing the front. Arrows, injury-dependent β-galactosidase expression. dpi, days post-injury. **b**, Whole-mount images of X-gal stained hearts and paws from *LEN-hsp68::lacZ* line 6, which exhibited vascular endothelial expression in uninjured hearts and paws. Scale bars represent 1 mm.



Extended Data Figure 7 | Transgenic assays examining *lepb*-linked regeneration enhancer fragments in combination with *lepb P2* (fin regeneration). a, Schematic representation of the transgenic constructs to examine *LEN* fragments that direct expression during fin regeneration. Expression during fin regeneration and the number of stable lines is indicated. **b**, Images of representative 0 dpa and 2 dpa fins from **a**. eGFP fluorescence is rarely detectable in uninjured fins. *LEN*(1–850), *LEN*(450–1000), *LEN*(450–850), and *LEN*(660–850) coupled with *P2* directed *eGFP* expression during fin regeneration. **LEN*(830–1350)*P2:eGFP* lines exhibited very weak eGFP expression in fin regenerates, detectable with long exposure times and at high magnification (data not shown), suggesting the possibility of minor fin regeneration enhancer elements in 850–1000. At least 5 fish from each transgenic line were examined, and all animals displayed a similar expression pattern. Arrowheads, amputation planes.

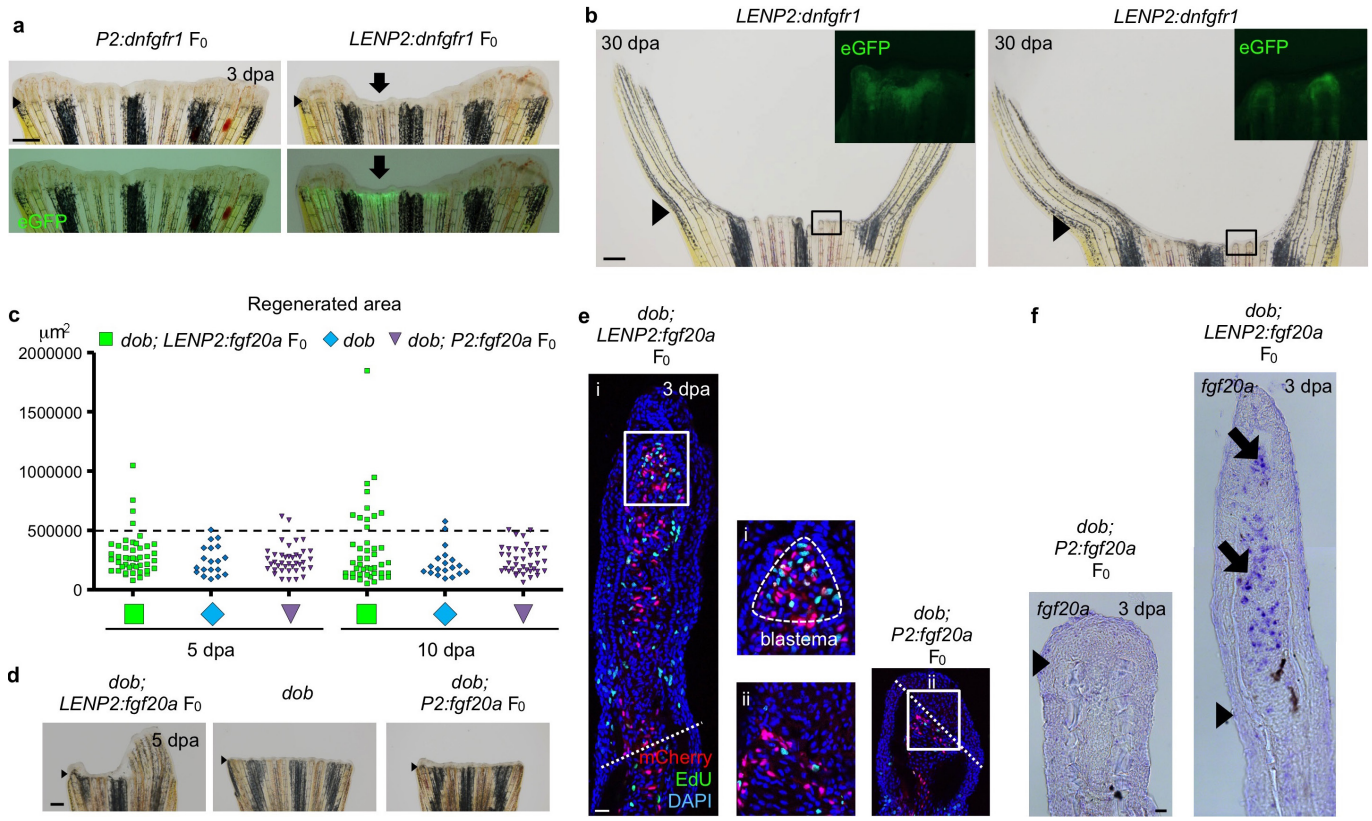


Extended Data Figure 8 | Images of heart sections from uninjured and regenerating transgenic lines that employ *lepb*-linked enhancer fragments. a–**h**, eGFP fluorescence is rarely detectable in uninjured hearts in all transgenic lines. One exception is *LEN*(1000–1350)*P2:eGFP*, which showed occasional, weak endocardial eGFP expression in uninjured hearts. *LEN*(1–850)*P2:eGFP* (**a**), *LEN*(450–1000)*P2:eGFP* (**b**), *LEN*(450–850)*P2:eGFP* (**d**), and *LEN*(660–850)*P2:eGFP* (**g**) transgenic lines, which include distal *LEN* elements, directed eGFP expression from promoters in a subset of epicardial cells and/or cardiomyocytes, but not endocardial cells. *LEN*(450–660)*P2:eGFP* lines (**e**) showed regeneration-dependent enhancer activity in cardiomyocytes near the injury site, but not in endocardial cells. Our data indicated that the activities of *LEN*(1–850)*P2:eGFP* (**a**), *LEN*(450–1000)*P2:eGFP* (**b**), and *LEN*(450–850)*P2:eGFP* (**d**) lines were not as strong as *LEN*(450–660)*P2:eGFP* (**e**), suggesting that there might be repressive elements for cardiomyocyte expression outside of sequences 450–660. *LEN*(830–1350) (**c**) and *LEN*(1000–1350) (**h**), which did not activate expression from promoters during fin regeneration, could direct endocardial expression in both ventricle and atrium during regeneration, similar to the reference reporters *lepb:eGFP* and *LENP2:eGFP*. Arrows in **c**, **h**, endocardial eGFP. i, ii, Enlargements of the boxed areas in regenerating ventricle and atrium, respectively. At least 5 fish from each transgenic line were examined, and all animals displayed a similar expression pattern. Scale bars represent 50 μm.



Extended Data Figure 9 | See next page for caption.

Extended Data Figure 9 | Transgenic assays to examine *lepb*-linked enhancer fragment activity in combination with *cmlc2* and α-*cry* promoters. a, Schematic representation of the transgenic constructs to examine enhancer fragment activity in combination with the *cmlc2* promoter. Expression during fin regeneration and the number of stable lines is indicated. **b**, Images of representative 0 dpa and 2 dpa fins from **a**. eGFP fluorescence was very weak or undetectable in 0 dpa or uninjured fins. (1–850), (450–1000), (450–850), and (660–850) *LEN* fragments coupled with the *cmlc2* promoter activated blastemal eGFP fluorescence (arrows) during fin regeneration. One *LEN*(1–850)*cmlc2:eGFP* line did not show fin regeneration enhancer activity. Arrowheads, amputation planes. At least five fish from each transgenic line were examined, and all animals displayed a similar expression pattern except for the following: For two strains of *LEN*(450–850)*cmlc2:eGFP*, 4 of 5 animals induced eGFP fluorescence at 2 dpa; For *LEN*(660–850)*cmlc2:eGFP*, 4 of 7 animals induced eGFP fluorescence at 2 dpa. **c**, Left: schematic diagram of the *LEN*(1000–1350)*cmlc2:eGFP* transgenic construct. Right: images of sections from uninjured and regenerating *LEN*(1000–1350)*cmlc2:eGFP* hearts. eGFP is expressed mosaically in cardiomyocytes via the *cmlc2* promoter. Uninjured hearts had no detectable endocardial eGFP fluorescence, whereas 3 dpa hearts displayed induced endocardial eGFP fluorescence (arrows). Arrowheads indicate cardiomyocyte eGFP fluorescence driven by *cmlc2* promoter activity. **d**–**h**, Schematic representation of the transgenic constructs to examine enhancer fragment activity in combination with the α-*cry* promoter. Expression during fin regeneration and injury-activated endocardial expression, and the number of stable lines are indicated. At least 5 fish from each transgenic line were examined, and all animals displayed a similar expression pattern. EC, endocardial cells. One *LEN*α-*cry:eGFP* line showed regeneration-dependent expression (arrows) in fins (**e**); yet, unlike when coupled with *lepb* and *cmlc2* promoters, the *LEN*(450–1000) fragment did not direct expression during fin regeneration (**d** and data not shown). This suggests a possible repressive motif within α-*cry* sequences. Asterisk indicates that one *LEN*α-*cry:eGFP* line showed weak endocardial eGFP expression in uninjured hearts, but the eGFP signal (arrows) was stronger and broader during regeneration (**g**). Two *LEN*(830–1350)α-*cry:eGFP* lines had no detectable eGFP fluorescence in regenerating fins (**f**) or uninjured hearts (**h**), but displayed induced endocardial eGFP fluorescence (arrows) during heart regeneration (**h**). i, ii, Enlargements of the boxed areas in regenerating ventricle and atrium, respectively. **i**, *LEN* sequences annotated with putative binding sites in fin (663–854) and cardiac (1034–1350) regeneration enhancer modules. **j**, A predicated AP-1 binding site is necessary for fin regeneration enhancer activity. Top, schematic representation of the *LEN(*450–850-*AP-1mut)P2* transgenic construct, in which the predicted AP-1 binding site (TGACTCA) is mutated to AAAAAA. Two *LEN*(450–850-*AP-1mut)P2* lines had no detectable eGFP fluorescence in regenerating fins. Scale bars represent 50 μm.



Extended Data Figure 10 | Pairing *LEN* with potent developmental influences can control regenerative capacity. a, Images of representative F_0 transgenic zebrafish injected with *P2:dnfgfr1* (left) or *LENP2:dnfgfr1* (right) constructs, shown at 3 dpa. The *dn-fgfr1* cassette is fused in frame to *eGFP*. Whereas zero of 27 *P2:dnfgfr1* F_0 animals displayed defective regeneration, 7 of 67 *LENP2:dnfgfr1* F_0 zebrafish had impaired fin regeneration in some fin rays, corresponding to eGFP fluorescence (arrow). **b**, Additional examples of *LENP2:dnfgfr1* fins at 30 dpa, from experiments with a stable line. Inset in **b**, high magnification view of the boxed area, showing eGFP fluorescence. **c**, Quantification of regenerated area from *dob*; *LENP2:fgf20a* F_0 transgenic zebrafish ($n = 45$, 44 at 5, 10 dpa, respectively), *dob* mutants ($n = 19$, 19 at 5, 10 dpa, respectively), and *dob; P2:fgf20a* F_0 transgenic zebrafish ($n = 40$, 40 at 5, 10 dpa, respectively) at 5 dpa and 10 dpa. Dotted line indicates 500,000 μm^2. **d**, Images of representative *dob; LENP2:fgf20a* F_0 transgenic zebrafish, *dob* mutants, and *dob; P2:fgf20a* F_0 transgenic zebrafish at 5 dpa. **e**, Confocal images of tissue sections of 3 dpa fin regenerates. Mosaic regenerates indicate expression of the linked *ef1α*:nls–mCherry marker construct (red), and EdU incorporation (collected 60 min after injection; green). DAPI, blue. F_0 mosaic *dob; LENP2:fgf20a* regenerates show evidence of distal growth and blastemal EdU incorporation. Arrow, blastema. Dotted lines, amputation planes. i, ii, Enlargements of the boxed areas. **f**, *In situ* hybridization in sections of 3 dpa fin regenerates from *dob; P2:fgf20a* (left) and F_0 mosaic *dob; LENP2:fgf20a* (right) animals, indicating *LEN*-induced *fgf20a* expression in mesenchymal cells and regenerative growth (arrows). *fgf20a* is rarely detected in *dob; P2:fgf20a* regenerates. Arrowheads, amputation planes.

LETTER

doi:10.1038/nature17169

A map of the large day–night temperature gradient of a super-Earth exoplanet

Brice-Olivier Demory[1], Michael Gillon[2], Julien de Wit[3], Nikku Madhusudhan[4], Emeline Bolmont[5], Kevin Heng[6], Tiffany Kataria[7], Nikole Lewis[8], Renyu Hu[9,10], Jessica Krick[11], Vlada Stamenković[9,10], Björn Benneke[10], Stephen Kane[12] & Didier Queloz[1]

Over the past decade, observations of giant exoplanets (Jupiter-size) have provided key insights into their atmospheres[1,2], but the properties of lower-mass exoplanets (sub-Neptune) remain largely unconstrained because of the challenges of observing small planets. Numerous efforts to observe the spectra of super-Earths—exoplanets with masses of one to ten times that of Earth—have so far revealed only featureless spectra[3]. Here we report a longitudinal thermal brightness map of the nearby transiting super-Earth 55 Cancri e (refs 4, 5) revealing highly asymmetric dayside thermal emission and a strong day–night temperature contrast. Dedicated space-based monitoring of the planet in the infrared revealed a modulation of the thermal flux as 55 Cancri e revolves around its star in a tidally locked configuration. These observations reveal a hot spot that is located 41 ± 12 degrees east of the substellar point (the point at which incident light from the star is perpendicular to the surface of the planet). From the orbital phase curve, we also constrain the nightside brightness temperature of the planet to 1,380 ± 400 kelvin and the temperature of the warmest hemisphere (centred on the hot spot) to be about 1,300 kelvin hotter (2,700 ± 270 kelvin) at a wavelength of 4.5 micrometres, which indicates inefficient heat redistribution from the dayside to the nightside. Our observations are consistent with either an optically thick atmosphere with heat recirculation confined to the planetary dayside, or a planet devoid of atmosphere with low-viscosity magma flows at the surface[6].

We observed the super-Earth 55 Cancri e for 75 h in total from 15 June to 15 July 2013 in the 4.5-μm channel of the Spitzer Space Telescope Infrared Array Camera (IRAC). The observations were split into eight continuous visits, each spanning 9 h and corresponding to half of the 18-h orbital period of 55 Cancri e. We acquired a total of 4,981,760 frames in subarray mode with an individual 0.02-s integration time. We extract the photometric time series from the raw frames using a previously described[4] aperture photometry pipeline. Each of the eight resulting light curves exhibit periodic flux variations due to the strong intra-pixel sensitivity of the IRAC detector combined to Spitzer's pointing wobble. The data reduction of this data set has been published elsewhere[7], but a summary is provided in Methods.

We analyse the light curves using a Markov chain Monte Carlo (MCMC) algorithm[8]. We simultaneously fit the eight half phase curves and a model of the detector systematics. Our MCMC algorithm includes an implementation of a pixel-level correction[9] and propagates the contribution from correlated noise in the data to the system best-fit parameters. In our implementation of the method, we build a sub-pixel mesh of n^2 grid points, evenly distributed along the x and y axes. Similar to a previous study[10], we find that the full-width at half-maximum (FWHM) of the point response function (PRF) along the x and y axes of the detector evolves with time and allows further improvement to the systematics correction. We thus combine the pixel-mapping algorithm with a linear function of the FWHM of the PRF along each axis. We find that this model provides the best correction to the data. The free planetary parameters in the MCMC fit are the phase-curve amplitude and offset (the angle between the peak of the modulation and the sub-stellar point), the occultation depth, the impact parameter, the orbital period, the transit centre and the transit depth. The functional form of the phase curve used in this fit is detailed in Methods. We combine the data points into 30-s bins for computing efficiency, which has previously been shown to have no effect on the derived parameters[7,11]. We find an average photometric precision of 363 p.p.m. per 30 s, and evaluate the level of correlated noise in the data for each data set using a time-averaging technique[12]. Results from the MCMC fits are shown in Table 1. We perform two additional analyses of this data set (see Methods) using a different model for the pixel-level correction, which results in phase-curve parameters in agreement with our main analysis.

Table 1 | 55 Cancri e planetary parameters

Basic planetary parameters	
Planet-to-star radius ratio, $R_p/R_\star$	$0.0187^{+0.0007}_{-0.0007}$
$b = a\cos(i)/R_\star$ ($R_\star$)	$0.41^{+0.05}_{-0.05}$
$T_0 - 2{,}450{,}000$ (BJD_{TDB})	$5733.013^{+0.007}_{-0.007}$
Orbital period, P (days)	$0.736539^{+0.000007}_{-0.000007}$
Orbital semi-major axis, a (AU)	$0.01544^{+0.00009}_{-0.00009}$
Orbital inclination, i (°)	$83.3^{+0.9}_{-0.8}$
Mass*, M_p ($M_\oplus$)	$8.08^{+0.31}_{-0.31}$
Radius, R_p ($R_\oplus$)	$1.91^{+0.08}_{-0.08}$
Mean density, ρ_p (g cm^{-3})	$6.4^{+0.8}_{-0.7}$
Surface gravity, log[g_p (cm s^{-2})]	$3.33^{+0.04}_{-0.04}$
Planetary emission parameters from this work	
Phase-curve amplitude, A_{phase} (p.p.m.)	197 ± 34
Phase-curve offset, θ_{phase} (degrees east)	41 ± 12
Mid-eclipse occultation depth (p.p.m.)	154 ± 23
Maximum hemisphere-averaged temperature (K)	$2{,}697^{+268}_{-275}$
Minimum hemisphere-averaged temperature (K)	$1{,}376^{+344}_{-451}$
Average dayside temperature (K)	$2{,}349^{+188}_{-193}$

Results are from the MCMC combined fit. Values indicated are the median of the posterior distributions and the 1σ credible intervals. $R_\star$, stellar radius; b, impact parameter; T_0, transit centre; BJD, barycentric Julian date; TDB, barycentric dynamical time; $M_\oplus$, Earth mass; $R_\oplus$, Earth radius. *Mass prior distribution obtained from ref. 30.

[1]Astrophysics Group, Cavendish Laboratory, JJ Thomson Avenue, Cambridge CB3 0HE, UK. [2]Institut d'Astrophysique et de Géophysique, Université of Liège, allée du 6 Aout 17, 4000 Liège, Belgium. [3]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA. [4]Institute of Astronomy, University of Cambridge, Cambridge CB3 0HA, UK. [5]NaXys, Department of Mathematics, University of Namur, 8 Rempart de la Vierge, 5000 Namur, Belgium. [6]University of Bern, Center for Space and Habitability, Sidlerstrasse 5, CH-3012, Bern, Switzerland. [7]Astrophysics Group, School of Physics, University of Exeter, Stocker Road, Exeter EX4 4QL, UK. [8]Space Telescope Science Institute, Baltimore, Maryland 21218, USA. [9]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA. [10]Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, California 91125, USA. [11]Spitzer Science Center, MS 220-6, California Institute of Technology, Jet Propulsion Laboratory, Pasadena, California 91125, USA. [12]Department of Physics and Astronomy, San Francisco State University, 1600 Holloway Avenue, San Francisco, California 94132, USA.

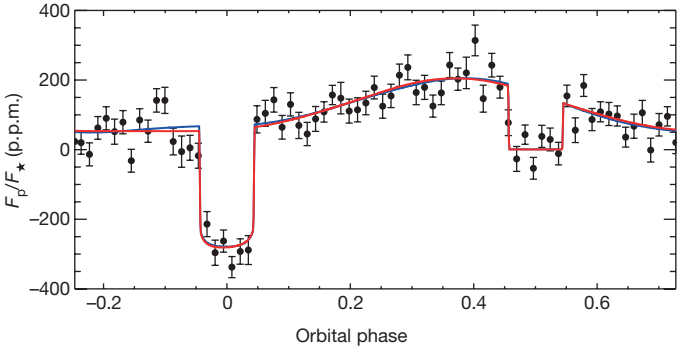


Figure 1 | 55 Cancri e Spitzer/IRAC 4.5-μm phase curve. Photometry for all eight data sets combined and folded onto the 0.74-day orbital period of 55 Cancri e. The black filled circles represent the relative flux ($F_p/F_\star$) variation in phase and are data binned per 15 min. The best-fit model using a three-longitudinal-band model is shown in red; the best-fit model using a one-longitudinal-band model is shown in blue. The error bars are the standard deviation of the mean within each orbital-phase bin.

The combined light curve (Fig. 1) exhibits a flux increase that starts slightly before the transit and reaches a maximum at 2.1 ± 0.6 h before opposition. We find a phase-curve peak amplitude of 197 ± 34 p.p.m., a minimum of 48 ± 34 p.p.m. and an occultation depth of 154 ± 23 p.p.m. (mid-eclipse).

We find that stellar variability could not cause the observed phase variation. The host is known to be an old, quiet star with a rotation period of 42 days that shows, on rare occasions, variability at the 6-millimagnitude level, corresponding to <1% coverage in star spots[13]. The periodic modulation that we observe is equal to the planetary orbital period and has a shape that remains consistent over the 4 weeks of the Spitzer observations. At infrared wavelengths, the effect of star-spots on the photometry is markedly reduced[14], but it is still possible that 1% spot coverage could produce a signal of the order of 200 p.p.m. However, the periodicity of the signal produced by such a starspot would be similar to the stellar rotation.

We also investigate the amplitude of the ellipsoidal effects[15] caused by 55 Cancri e on its host star and find an expected amplitude of 0.6 p.p.m. The reciprocal effect from the host star on the planet would translate to an effect of about 1 p.p.m. (ref. 16). None of these features would be detectable in our data set. In addition, ellipsoidal variations have a frequency that is twice that of the orbital period of the planet. For these two reasons, we discard the possibility that ellipsoidal variations are at the origin of the observed signal.

An alternative way to mimic the orbital phase curve would be a scenario in which 55 Cancri e induces starspots on the stellar surface via magnetic field interactions, which would produce a photometric modulation that is synchronized with the orbital period of the planet[17]. It has been suggested that the amplitude of these interactions increases with the ratio of the planetary mass to its orbital semi-major axis[17]; however, currently, there is no robust evidence for star–planet interactions even for planets with masses of 3–5 that of Jupiter on 0.9–5-day orbital periods. 55 Cancri e is an exoplanet with a mass of 0.02 Jupiter masses in a 0.74-day orbit; considering the large body of work on star–planet interactions[18], we deem it unlikely that 55 Cancri e could induce synchronized starspot patterns on its host star. Therefore, we assume in the following that the observed modulation originates from the planet itself.

The shape of the phase curve of 55 Cancri e provides constraints on the thermal brightness map of the planet. The phase-curve amplitude translates to a warmest-hemisphere-averaged brightness temperature of $2{,}697^{+268}_{-275}$ K at 4.5 μm, and a coolest-hemisphere-averaged brightness temperature of $1{,}376^{+344}_{-451}$ K. We find that the hot spot is centred on the meridian located $41° \pm 12°$ east of the substellar point. We longitudinally map the dayside of 55 Cancri e using an MCMC implementation[19]. This method was developed to map exoplanets and to mitigate the degeneracy between the planetary brightness distribution and the system parameters. We model the planetary dayside using two different prescriptions, similar to a previous study[20]. In the first model, we use a single longitudinal band (Fig. 2, left) with a position and width that are adjusted in the MCMC fit. The second model is similar to the 'beach-ball model'[21] that uses three longitudinal bands with fixed positions and widths (Fig. 2, right). In both cases, the relative brightness between each longitudinal band is adjusted in the MCMC fit.

The large day–night temperature difference of more than 1,300 K indicates a lack of strong atmospheric circulation redistributing energy from the dayside to the nightside of the planet. Such a large contrast could potentially be explained by the extremely high stellar irradiation received on the dayside, owing to which the radiative timescale might be shorter than the advective timescale, as has been suggested for highly irradiated hot Jupiters, which have H_2-dominated atmospheres[22]. However, the mass, radius and temperature of the planet are inconsistent with the presence of an H_2-dominated atmosphere[7,23], which is supported by the non-detection of H absorption in the Lyman-α[24] region of the spectrum, although an atmosphere with a higher mean molecular weight cannot be ruled out. It is possible that a high-mean-molecular-weight atmosphere of 55 Cancri e, for example, consisting largely of H_2O or CO_2, could also have a lower radiative timescale compared to the advective timescale, thereby explaining the inefficient circulation. However, the observed brightness temperature is unexpectedly high for such an explanation, because H_2O and CO_2 both have substantial opacity in the IRAC 4.5-μm bandpass, owing to which the upper, cooler regions of the atmosphere are probed preferentially.

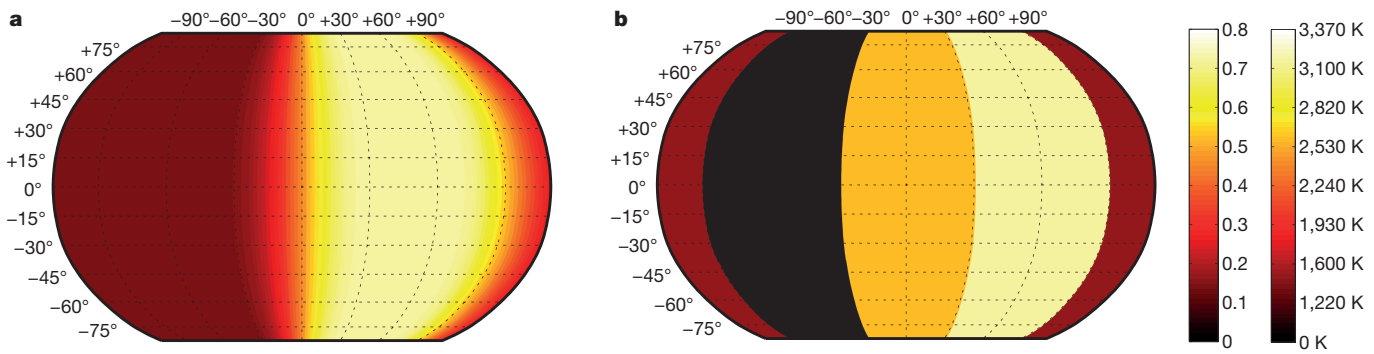


Figure 2 | Longitudinal brightness maps of 55 Cancri e. Longitudinal brightness distributions as retrieved from the Spitzer/IRAC 4.5-μm phase curve. The planetary dayside is modelled using two prescriptions. **a**, One-longitudinal-band model, with the band position, width and brightness adjusted in the fit. **b**, Three-longitudinal-band model, with the band positions and widths fixed, but their relative brightnesses adjustable. The colour scales indicate the planetary brightness normalized to the stellar-average brightness and the brightness temperature for each longitudinal band.

The maximum hemisphere-averaged temperature of $2,700 \pm 270$ K is marginally greater than the highest permissible equilibrium temperature, which is possible for the planetary surface, but implausible higher up in the atmosphere unless the atmosphere is host to strong thermal inversion[25]. Alternatively, the data may be explained by a planet devoid of a thick atmosphere of any composition that also has a low albedo. Such a hypothesis could explain both the radius of the planet, which is consistent with a purely rocky composition, as well as the lack of strong atmospheric circulation.

The substantial day–night temperature contrast observed is seemingly incongruous with the observation of a large offset of the hot spot 41° east from the substellar point. Such a shift of the hot spot requires efficient energy circulation in the atmosphere[22], contrary to the large day–night contrast observed. An alternative explanation is that the planet harbours an optically thick atmosphere in which heat recirculation takes place, but only on the dayside, while the gases condense out on the planetary nightside[26], possibly forming clouds[20]. However, such a scenario requires either the atmosphere to be dominated by vapours of high-temperature refractory compounds, for example, silicates[27,28], or the nightside temperatures to be below freezing so that volatiles such as H_2O condense; the latter is ruled out by our observed nightside temperature of $1,380 \pm 400$ K. It is possible that there are strong longitudinal inhomogeneities in the chemical composition and emissivity in the atmosphere causing a longitudinally varying optical depth in the 4.5-μm bandpass that could potentially explain the data. Alternatively, the hot-spot offset may be driven by an eastward molten lava flow on the dayside surface of the planet, which would have a viscosity more similar to water at room temperature than to solid rock. At the observed maximum hemisphere-averaged temperature of about 2,700 K, silicate-based rocks are expected to be molten[29], whereas the nightside temperature of about 1,380 K can be cool enough to sustain a partially to mostly solid surface, where rock viscosities would be several orders of magnitude larger than on the dayside.

Additional constraints resulting from the estimated atmospheric escape induced by the nearby host star suggest that it is unlikely that 55 Cancri e is harbouring a thick atmosphere. We find that the surface pressure of 55 Cancri e needs to be larger than 31 kbar for the planet to survive over the lifetime of its star, which supports a scenario in which 55 Cancri e has no atmosphere (see Methods).

From the fit from the three-longitudinal-band model, we find that the region of maximum thermal emission is located 30°–60° east of the substellar point, with brightness temperatures in excess of 3,100 K. We find that tidal dissipation can explain only a fraction of this re-emitted radiation (see Methods), suggesting that an additional, currently unknown source provides a sizeable contribution to the infrared emission of 55 Cancri e.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 27 August 2015; accepted 21 January 2016.
Published online 30 March 2016.

1. Burrows, A. S. Highlights in the study of exoplanet atmospheres. *Nature* **513,** 345–352 (2014).
2. Heng, K. & Showman, A. P. Atmospheric dynamics of hot exoplanets. *Annu. Rev. Earth Planet. Sci.* **43,** 509–540 (2015).
3. Knutson, H. A. *et al. Hubble Space Telescope* near-IR transmission spectroscopy of the super-Earth HD 97658b. *Astrophys. J.* **794,** 155 (2014).
4. Demory, B.-O. *et al.* Detection of a transit of the super-Earth 55 Cancri e with warm *Spitzer*. *Astron. Astrophys.* **533,** A114 (2011).
5. Winn, J. N. *et al.* A super-Earth transiting a naked-eye star. *Astrophys. J.* **737,** L18 (2011).
6. Solomatov, V. in *Treatise on Geophysics* Vol. 9 (ed. Schubert, G.) 91–119 (Elsevier, 2007).
7. Demory, B.-O., Gillon, M., Madhusudhan, N. & Queloz, D. Variability in the super-Earth 55 Cnc e. *Mon. Not. R. Astron. Soc.* **455,** 2018–2027 (2016).
8. Gillon, M. *et al.* The TRAPPIST survey of southern transiting planets. I. Thirty eclipses of the ultra-short period planet WASP-43 b. *Astron. Astrophys.* **542,** A4 (2012).
9. Stevenson, K. B. *et al.* Transit and eclipse analyses of the exoplanet HD 149026b using BLISS mapping. *Astrophys. J.* **754,** 136 (2012).
10. Lanotte, A. A. *et al.* A global analysis of *Spitzer* and new HARPS data confirms the loneliness and metal-richness of GJ 436 b. *Astron. Astrophys.* **572,** A73 (2014).
11. Deming, D. *et al. Spitzer* secondary eclipses of the dense, modestly-irradiated, giant exoplanet HAT-P-20b using pixel-level decorrelation. *Astrophys. J.* **805,** 132 (2015).
12. Pont, F., Zucker, S. & Queloz, D. The effect of red noise on planetary transit detection. *Mon. Not. R. Astron. Soc.* **373,** 231–242 (2006).
13. Fischer, D. A. *et al.* Five planets orbiting 55 Cancri. *Astrophys. J.* **675,** 790–801 (2008).
14. Berta, Z. K. *et al.* The GJ1214 super-Earth system: stellar variability, new transits, and a search for additional planets. *Astrophys. J.* **736,** 12 (2011).
15. Mazeh, T. & Faigler, S. Detection of the ellipsoidal and the relativistic beaming effects in the CoRoT-3 lightcurve. *Astron. Astrophys.* **521,** L59 (2010).
16. Budaj, J. The reflection effect in interacting binaries or in planet–star systems. *Astron. J.* **141,** 59 (2011).
17. Shkolnik, E., Bohlender, D. A., Walker, G. A. H. & Collier Cameron, A. The on/off nature of star–planet interactions. *Astrophys. J.* **676,** 628–638 (2008).
18. Miller, B. P., Gallo, E., Wright, J. T. & Pearson, E. G. A comprehensive statistical assessment of star–planet interaction. *Astrophys. J.* **799,** 163 (2015).
19. de Wit, J., Gillon, M., Demory, B.-O. & Seager, S. Towards consistent mapping of distant worlds: secondary-eclipse scanning of the exoplanet HD 189733b. *Astron. Astrophys.* **548,** A128 (2012).
20. Demory, B.-O. *et al.* Inference of inhomogeneous clouds in an exoplanet atmosphere. *Astrophys. J.* **776,** L25 (2013).
21. Cowan, N. B. *et al.* Alien maps of an ocean-bearing world. *Astrophys. J.* **700,** 915–923 (2009).
22. Showman, A. P., Fortney, J. J., Lewis, N. K. & Shabram, M. Doppler signatures of the atmospheric circulation on hot Jupiters. *Astrophys. J.* **762,** 24 (2013).
23. Gillon, M. *et al.* Improved precision on the radius of the nearby super-Earth 55 Cnc e. *Astron. Astrophys.* **539,** A28 (2012).
24. Ehrenreich, D. *et al.* Hint of a transiting extended atmosphere on 55 Cancri b. *Astron. Astrophys.* **547,** A18 (2012).
25. Madhusudhan, N. & Seager, S. On the inference of thermal inversions in hot Jupiter atmospheres. *Astrophys. J.* **725,** 261–274 (2010).
26. Heng, K. & Kopparla, P. On the stability of super-Earth atmospheres. *Astrophys. J.* **754,** 60 (2012).
27. Schaefer, L. & Fegley, B. Jr. Atmospheric chemistry of Venus-like exoplanets. *Astrophys. J.* **729,** 6 (2011).
28. Miguel, Y., Kaltenegger, L., Fegley, B. & Schaefer, L. Compositions of hot super-Earth atmospheres: exploring *Kepler* candidates. *Astrophys. J.* **742,** L19 (2011).
29. Lutgens, F. K. & Tarbuck, E. J. *Essentials of Geology* 7th edn, Ch. 3 (Prentice Hall, 2000).
30. Nelson, B. E. *et al.* The 55 Cancri planetary system: fully self-consistent *N*-body constraints and a dynamical analysis. *Mon. Not. R. Astron. Soc.* **441,** 442–451 (2014).

Acknowledgements We thank D. Deming, D. Apai and A. Showman for discussions as well as the Spitzer Science Center staff for their assistance in the planning and executing of these observations. This work is based on observations made with the Spitzer Space Telescope, which is operated by the Jet Propulsion Laboratory, California Institute of Technology under a contract with NASA. Support for this work was provided by NASA through an award issued by JPL/Caltech. M.G. is a Research Associate at the Belgian Funds for Scientific Research (FRS-FNRS). V.S. was supported by the Simons Foundation (award number 338555, VS).

Author Contributions B.-O.D. initiated and led the Spitzer observing programme, conducted the data analysis and wrote the paper. M.G. performed an independent analysis of the dataset. E.B. carried out the simulations assessing the amplitude of tidal heating in the interior of 55 Cancri e. J.d.W. performed the longitudinal mapping of the planet. N.M. wrote the interpretation section with inputs from E.B., K.H., V.S., R.H., N.L. and T.K. J.K., B.B., S.K. and D.Q. contributed to the observing programme. All authors commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.-O.D. (bod21@cam.ac.uk).

METHODS

Data reduction. We observed four phase curves of 55 Cancri e with the Spitzer Space Telescope in the IRAC/4.5-μm channel as part of our program ID 90208. Because of downlink constraints, these four phase-curve observations were split into eight separate observations (or Astronomical Observation Requests, AORs) each lasting half an orbit of 55 Cancri e. Details of each AOR are provided in Extended Data Table 1. The corresponding data can be accessed from the Spitzer Heritage Archive (http://sha.ipac.caltech.edu). All AORs were acquired in stare mode using a constant exposure time (0.02 s). All our data were obtained using the Pointing Calibration and Reference Sensor (PCRS) peak-up mode, which allows the observer to place the target on a precise location on the detector to mitigate the intra-pixel sensitivity variations. This observing mode increases the pointing stability and reduces the level of correlated noise in the data by a factor of 2–3 (ref. 31). AOR 48072960 experienced a 30-min interruption during data acquisition, which forces us to treat both parts of that AOR separately in the rest of this section. We do not retain the 30-min-long PCRS sequences in our analysis because the motion of the star on the detector yields large correlated noise in these data sets. Our reduction uses the basic calibrated data (BCD) that are downloaded from the Spitzer Archive. The BCD are Flexible Image Transfer System (FITS) data cubes consisting of 64 frames of 32×32 pixels each. Our data reduction code reads each frame, converts fluxes from the Spitzer units of specific intensity (MJy sr^{-1}) to photon counts, and transforms the data time-stamps from $\mathrm{BJD_{UTC}}$ to $\mathrm{BJD_{TDB}}$ using existing procedures[32]. We did not deem it necessary to discard specific sub-array frames. During the reduction process, we compute the flux, position and FWHM in each of the 64 frames of each data cube; the frames for which any of these parameters differ from the median by more than 5σ are discarded. The centroid position on the detector is determined by fitting a Gaussian to the marginal x, y distributions using the *GCNTRD* procedure of the IDL Astronomy User's Library[33]. We also fit a two-dimensional Gaussian to the stellar PRF following previous studies[34]. We find that determining the centroid position using *GCNTRD* results in a smaller dispersion of the fitted residuals by 10% to 15% across our data set, in agreement with other Spitzer analyses[35]. We then perform aperture photometry for each data set using a modified version of the *APER* procedure using aperture sizes of ranging from 2.2 to 4.4 pixels in 0.2-pixel intervals. We choose the optimal aperture size on the basis of minimizing r.m.s. $\times\ \beta_{\mathrm{red}}^2$ for each AOR, where r.m.s. is the root-mean-square of the photometric time series and β_{red} is the red-noise contribution[8]. The red noise is assessed over 60-min timescales because shorter timescales are irrelevant for the phase-curve signal whose periodicity is 18 h. We measure the background contribution on each frame using an annulus located 10 to 14 pixels from the centroid position. Our code also determines the FWHM of the PRF along the x and y axes. We use a moving average based on forty consecutive frames to discard data points that differ from the median by more than 5σ in background, (x, y) position or FWHM. We find that, on average, 0.06% of the data points are discarded. The resulting time series are combined into 30-s bins to speed up the analysis; this binning has been shown to have no influence on the values or uncertainties of the system parameters[7]. We show the optimal aperture size, corresponding r.m.s. and β_{red} for each data set in Extended Data Table 1.

Photometric analysis. *Intra-pixel sensitivity correction.* We use an implementation of the BLISS (BiLinearly-Interpolated Sub-pixel Sensitivity)[9] to account for the intra-pixel sensitivity variations, as was similarly used in a previous study using the same data set[7].

The BLISS algorithm uses a bilinear interpolation of the measured fluxes to build a pixel-sensitivity map. The data are thus self-calibrated. Our implementation of this algorithm is included in the Markov chain Monte Carlo (MCMC) framework presented in ref. 8. The improvement introduced by a pixel-mapping technique such as BLISS requires that the stellar centroid remains in a relatively confined area on the detector, which warrants an efficient sampling of the x–y region and, hence, an accurate pixel map. In our implementation of the method, we build a sub-pixel mesh of n^2 grid points, evenly distributed along the x and y axes. The BLISS algorithm is applied at each step of the MCMC fit. The number of grid points is determined at the beginning of the MCMC by ensuring that at least five valid photometric measurements are located in each mesh box. Similar to two recent studies[7,10], we find that a further reduction of the level of correlated noise in the photometry is achieved by the inclusion of the FWHM of the PRF along the x and y axes as extra parameters in the baseline model. The PRF evolves with time and its properties are not accounted for by the BLISS algorithm. We thus combine the BLISS algorithm with a linear function of the FWHM of the PRF along the x and y axes. In addition, the baseline model for each AOR includes a flux constant. We find that including a model of the FWHM of the PRF decreases the Bayesian Information Criterion (BIC)[36] by $\Delta\mathrm{BIC} = 591$. We show the raw data sets with the best-fit instrumental + astrophysical model superimposed in red in Extended Data Figs 1–3. The corrected photometry is shown in Extended Data Figs 4–6. The phase-curve modulation is clearly noticeable in each AOR. The behaviour of the photometric r.m.s. as a function of binning is shown for each data set in Extended Data Fig. 7.

Model comparison. In our first MCMC analysis, to model the variation in the infrared emission of the planet we use $F = F_{\mathrm{p}} + \mathrm{Tr} + \mathrm{Oc}$ (in which F is the observed flux, F_{p} is the phase modulation driven by the planet, Tr is the transit model and Oc is the occulation model), and a Lambertian[37] functional form for F_{p}

$$F_{\mathrm{p}} = A_{\mathrm{phase}} \frac{\sin(z) + (\pi - z)\cos(z)}{\pi}$$

in which A_{phase} is the phase amplitude and

$$\cos(z) = -\sin(i)\cos[2\pi(\phi + \theta_{\mathrm{phase}})]$$
$$\phi = \frac{2\pi}{P}(t - T_0)$$

where θ_{phase} is the phase-curve offset, ϕ is the phase angle, i is the orbital inclination of the planet, P the orbital period, T_0 the transit centre and t is time.

The transit- (Tr) and occultation- (Oc) light-curve model MA (ref. 38) are summarized as

$$\mathrm{Tr} = \mathrm{MA}(d_{F_{\mathrm{tr}}}, P, b, M_\star, c_1, c_2, t)$$
$$\mathrm{Oc} = \mathrm{MA}(d_{F_{\mathrm{oc}}}, P, b, M_\star, t)$$

in which $d_{F_{\mathrm{tr}}}$ and $d_{F_{\mathrm{oc}}}$ are the transit and occultation depths, respectively, b is the impact parameter, $M_\star$ is the stellar mass, and $c_1 = 2u_1 + u_2$ and $c_2 = u_1 - 2u_2$ are the limb-darkening linear combinations, with u_1 and u_2 the quadratic coefficients obtained from theoretical tables[39] using published stellar parameters[40].

We also experimented using a sinusoid for the phase variation: $F_{\mathrm{p}} = A_{\mathrm{phase}}\cos(\phi + \theta_{\mathrm{phase}})$. The fit using a sinusoid results in an amplitude $A_{\mathrm{phase}} = 218 \pm 50$ p.p.m. and an offset value $\theta_{\mathrm{phase}} = 68° \pm 24°$ east of the substellar point, in agreement with our results using a Lambertian functional form ($A_{\mathrm{phase}} = 197 \pm 34$ p.p.m. and $\theta_{\mathrm{phase}} = 41° \pm 12°$).

The Lambertian sphere model provides a better fit to the data than does the sinusoid model, with $\Delta\mathrm{BIC} = 11$.

We also perform another MCMC analysis with no phase-curve model, hence removing two degrees of freedom (phase amplitude and phase offset). We find $\Delta\mathrm{BIC} = 21$ in favour of the model including the phase-curve model.

We also run an MCMC fit that includes the phase amplitude, but not the phase offset. We find that this fit produces only a marginal χ^2 improvement over the MCMC fit with the no-phase-curve model, but this improvement is penalized by the extra degree of freedom according to the BIC. We indeed obtain a $\Delta\mathrm{BIC} = 25$ in favour of the model including the phase-curve offset.

Altogether, this model comparison confirms that a phase-curve model that includes a phase offset is the favoured functional form according to the BIC.

Additional analyses. We conduct two additional analyses of our entire data set to assess the robustness of our initial detection that used the BLISS mapping technique. In these two analyses, we use different approaches to (1) model the intra-pixel sensitivity of the detector and (2) change the input data format.

In the first analysis, we use a simple polynomial detrending approach with a functional form that includes only the centroid position (fourth-order) and FWHM (first-order). We experimented with different polynomial orders (from one to four) for these two parameters and found that this combination globally minimizes the BIC. Each AOR has its own set of baseline coefficients. As for the BLISS mapping, the polynomial detrending is included in the MCMC fit so the baseline model and the system parameters are adjusted simultaneously to efficiently propagate the uncertainties to the final parameters. We find a level of correlated noise in the data that is only slightly larger (about 10%) than that obtained with the BLISS mapping technique. Using this method we find a phase-curve minimum of 36 ± 41 p.p.m., a maximum of 187 ± 41 p.p.m. and an offset of $50° \pm 13°$ east of the substellar point; using the BLISS mapping, the corresponding values are 47 ± 34 p.p.m., 197 ± 34 p.p.m. and $41° \pm 12°$ east. As previously shown[10], the addition of the FWHM of the PRF in the baseline model substantially improves a fit based on only a centroid position, and, most importantly, it enables an acceptable fit to 8-h time series.

In the second analysis, we aim to assess whether the phase-curve signal persists when we split our input data. All our AORs have durations of nearly 9 h, and we elect to split each of them in two to reduce the duration of each individual data set to 4.5 h. The functional forms of the baseline models are the same as for the analysis using the unsegmented input data, described above. In this additional test, we find a phase-curve minimum of 51 ± 51 p.p.m., a maximum of 216 ± 51 p.p.m. and an offset of $54° \pm 16°$ east of the substellar point. These results are in good

agreement with our main analysis. The uncertainties in the phase-curve parameters are larger in this case because of the time-series segmentation, which does not constrain the baseline coefficients as effectively as for longer data sets. The phase curves obtained from these additional analyses are shown in Extended Data Fig. 8.

We finally note that the phase-curve peak is located close to the start of half of our observations and towards the end of the other half data sets, which was necessary owing to Spitzer downlink limitations. We deem this pattern purely coincidental for two reasons. First, if our reported phase curve was due to uncorrected systematics, then it would be unlikely that the systematics would produce an upward trend in half of the data and a downward trend in the other half. These data sets are independent and there is no relationship between those obtained from transit to occultation and those obtained from occultation to transit. There is also no correlation with the centroid position on the detector. Second, if the phase-peak offset occurred after or before this discontinuity, then it would have been clearly detected in the continuous parts of our data set; however, only gradual slopes are seen in both data sets. A comparison with data obtained in the same year with the Microvariability and Oscillations of STars (MOST) satellite (D. Dragomir, personal communication) shows an agreement in the phase-curve amplitude and offset values derived from both facilities.

Longitudinal mapping. The key features of the phase curve of 55 Cancri e translate directly into constraints on maps[41,42] assuming a tidally locked planet on a circular orbit. A planetary phase curve $F_p/F_\star$ ($F_\star$ is the flux from the star) measures the planetary hemisphere-averaged relative brightness $\langle I_p\rangle/\langle I_\star\rangle$ as

$$\frac{F_p}{F_\star}(\alpha)=\frac{\langle I_p\rangle(\alpha)}{\langle I_\star\rangle}\left(\frac{R_p}{R_\star}\right)^2$$

in which α is the orbital phase, R_p is the planetary radius and $R_\star$ is the stellar radius.

The longitudinal mapping technique used here[19] aims to mitigate the degeneracy between the distribution of the planetary thermal brightness and the system parameters. This part of the analysis is independent of the light-curve analysis presented above. Therefore, here we fix the system parameters to those derived from a previous study[7], which is based on the entire 55 Cancri e Spitzer data set. Using this prior information for the purpose of longitudinal mapping is adequate because the degeneracy between the planetary brightness distribution and the system parameters is only relevant in the context of eclipse mapping[19]. We follow the same approach as for Kepler-7b (ref. 20) and use two families of models, similar to the 'beach-ball models' introduced in ref. 21: one using n longitudinal bands with fixed positions on the dayside, and another using longitudinal bands whose positions and widths are jump parameters in the MCMC fit. We choose a three-fixed-band model and one-free-band model to extract both the longitudinal dependence of the dayside brightness of 55 Cancri e and the extent of its 'bright' area. Increasing n to five yields a larger BIC than for $n=3$. For both models, we compute the amplitude of each band from their simulated light curve using a perturbed singular-value decomposition method. The one-free-band model (Fig. 2, left) yields a uniformly bright longitudinal area extending from $5°\pm18°$ west to $85°\pm18°$ east with a relative brightness of 0.72 ± 0.18, compared to a brightness of 0.15 ± 0.05 for the rest of the planet. The three-fixed-band model yields bands of relative brightness decreasing from the west to the east: <0.21 (3σ upper limit), 0.58 ± 0.15 and 0.74 ± 0.15, compared to the nightside contribution of 0.17 ± 0.06.

Variability of the thermal emission of 55 Cancri e. Variability in the thermal emission of 55 Cancri e between 2012 and 2013, has previously been determined from occultation measurements[7]. Several tests regarding the robustness of the variability pattern were conducted, including three different analyses that used BLISS mapping, polynomial detrending and a pixel-level decorrelation method[11]. These three approaches confirmed the variability of the thermal emission of the planet between 2012 and 2013 with similar uncertainties. Therefore, we consider it very likely that the emission of the planet is varying, but on timescales that are substantially longer than the timescale of the 2013 observations alone (a month) used here. No variability is reported in the 2013 data alone[7]. These factors justify our combining of the 2013 observations and our use of a single phase-curve model. Furthermore, we detect the phase-curve shape in all individual data sets in addition to the combined phase-folded time series. This strengthens our conclusion that it is unlikely that stellar variability would cause the combined phase-curve shape from individual stellar events taken at different times over the month of observations.

Brightness temperatures. We use an observed infrared spectrum of 55 Cancri e (ref. 43) to compute the brightness temperatures in the IRAC 4.5-μm bandpass from the $F_p/F_\star$ values derived from the MCMC fits.

Constraints on the atmosphere of 55 Cancri e. If an atmosphere was present, then the large temperature contrast between the dayside and nightside hemispheres suggests that the radiative cooling time (t_{rad}) is less than the dynamical time scale (t_{dyn}), resulting in a poor redistribution of heat from the dayside to the nightside. This sets a constraint on the mean molecular weight, which we may estimate. The zonal velocity is $v\approx\sqrt{\mathcal{R}\Delta T}\approx1$ km s^{-1}, in which $\mathcal{R}$ is the specific gas constant, $\Delta T=1{,}460$ K is the temperature difference between the hemispheres and we have ignored an order-unity correction factor associated with the pressure difference between the hemispheres[44]. If we enforce $t_{rad}<t_{dyn}$, then for the mean molecular weight μ we obtain

$$\mu>\frac{\mathcal{R}_{univ}(\Delta T)^{1/3}}{T_{day}^2}\left(\frac{P_{day}}{\sigma_{SB}R_p\kappa g}\right)^{2/3}$$

in which $\mathcal{R}_{univ}=8.3144598\times10^7$ erg K^{-1} g^{-1} is the universal gas constant, $T_{day}=2{,}700$ K is the dayside temperature, σ_{SB} is the Stefan–Boltzmann constant, $R_p=1.91R_\oplus$ is the planetary radius ($R_\oplus$ is the Earth radius), $\kappa=2/7$ is the adiabatic coefficient, $g=10^{3.33}$ cm s^{-2} is the surface gravity and P_{day} is the dayside pressure—the only unknown parameter in this expression. If we set $P_{day}=1$ bar, then $\mu>9$. This estimate further suggests that a hydrogen-dominated atmosphere is unlikely, and sets a lower limit on the mean molecular weight.

It is unlikely that 55 Cancri e is harbouring a thick atmosphere, owing to its proximity to its star. If we assume energy-limited escape[45], then the atmosphere needs to have sufficient mass to survive for the stellar age, which translates into a lower limit on the required surface pressure

$$P>\frac{\mathcal{L}_X R_p t_\star g}{16\pi G M_p a^2}$$

in which $\mathcal{L}_X$ is the X-ray luminosity of the star, $t_\star$ is the stellar age, G is Newton's gravitational constant, $M_p=8.08M_\oplus$ is the planetary mass ($M_\oplus$ is the Earth mass) and $a=0.01544$ AU is the orbital semi-major axis. If $\mathcal{L}_X=4\times10^{26}$ erg s^{-1} (ref. 24) and $t_\star=8$ Gyr, then $P>31$ kbar; in other words, the surface pressure of 55 Cancri e needs to be larger than 31 kbar to survive atmospheric escape over the stellar lifetime. Despite the uncertainties associated with estimating the mass loss due to atmospheric escape, this estimate is conservative because the star probably emitted higher X-ray luminosities in the past. Our suggestion of an atmosphereless 55 Cancri e is consistent with the trends predicted for super-Earths[45].

Hence, it is unlikely that the large infrared peak offset is due to an atmosphere rich with volatiles. It is more likely that the infrared phase curve of 55 Cancri e is probing non-uniformities associated with its molten rocky surface.

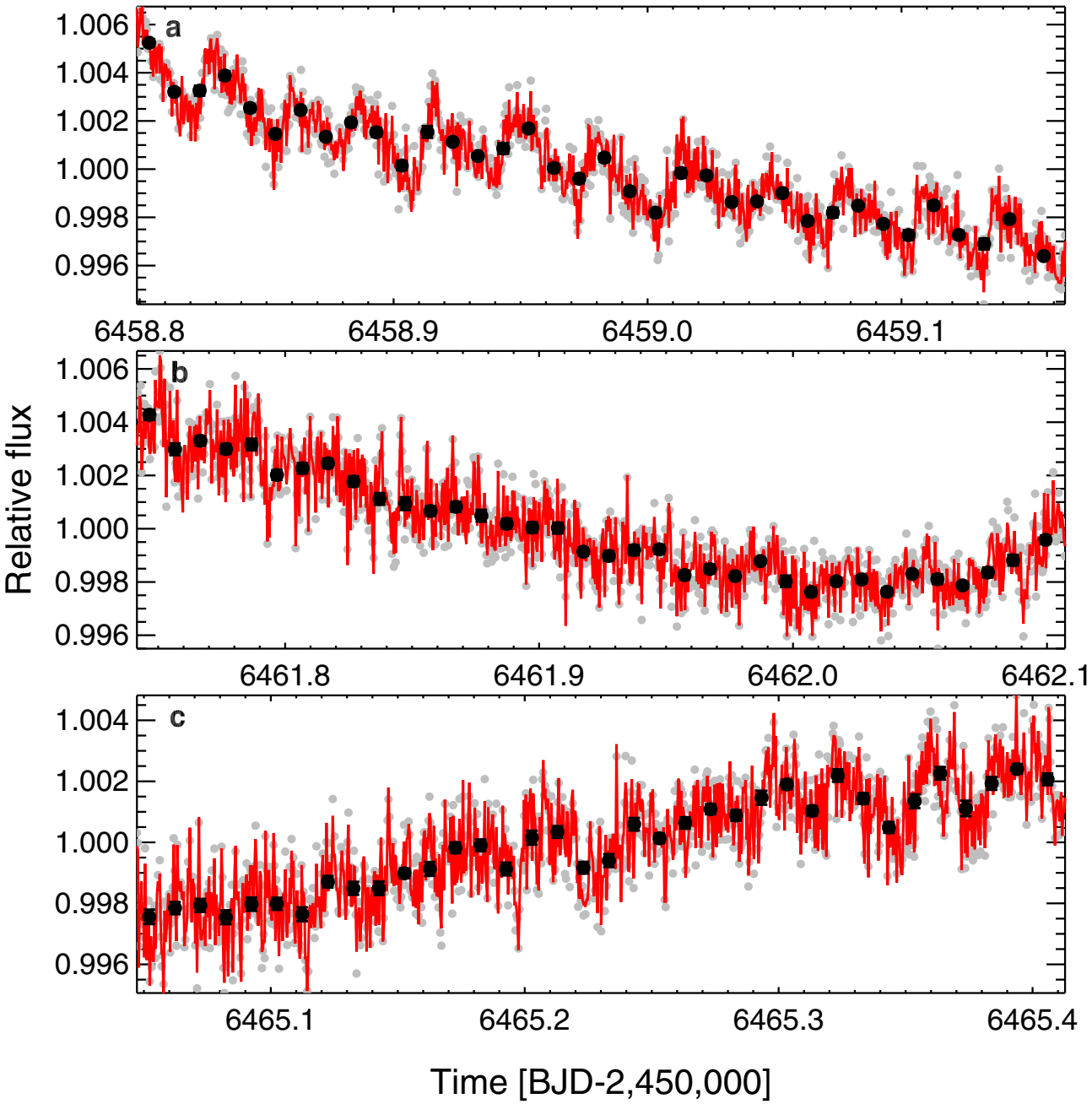
Tidal heating. Because 55 Cancri is a multi-planet system, the eccentricity and obliquity of 55 Cancri e are excited, owing to the presence of the outer planets. This creates a tidal heat flux that is responsible, in part, for the thermal emission of the planet. To evaluate the contribution of the tidal heat flux to the measured thermal emission, we investigate the possible values of the eccentricity and obliquity of 55 Cancri e for different tidal dissipation using N-body simulations (using Mercury-T; ref. 46). We use the orbital elements and masses for the four outer planets[30] and the most recent values[7] for the mass, radius and orbital semi-major axis of 55 Cancri e.

We find that the obliquity of 55 Cancri e is very low ($<1°$) and that the eccentricity is about 10^{-3} for the eight orders of magnitude (10^{-5}–10 times the dissipation of Earth $\sigma_\oplus$) we consider for the tidal dissipation of 55 Cancri e. The corresponding tidal heat flux ϕ_{tides} or tidal temperature $(\phi_{tides}/\sigma)^{1/4}$ increase with the dissipation in 55 Cancri e, from 10^{-3} W m^{-2} (a few kelvin) to 10^6 W m^{-2} (about 2,000 K). We calculate the occultation depth at 4.5-μm for a range of eccentricities and albedos (0.0–1.0) to enable a comparison with the output of the dynamical simulations (Extended Data Fig. 9). We find that a combination of large dissipation ($10\sigma_\oplus$), eccentricity and obliquity can explain the level of thermal emission observed in 2013; however, these solutions do not allow us to reproduce the nightside temperature. In our configuration (no heat re-distribution and assuming an isotropic tidal heat flux), tides do not match our measurements, so an additional heat source is probably responsible for at least part of the large planetary thermal emission observed in 2013.

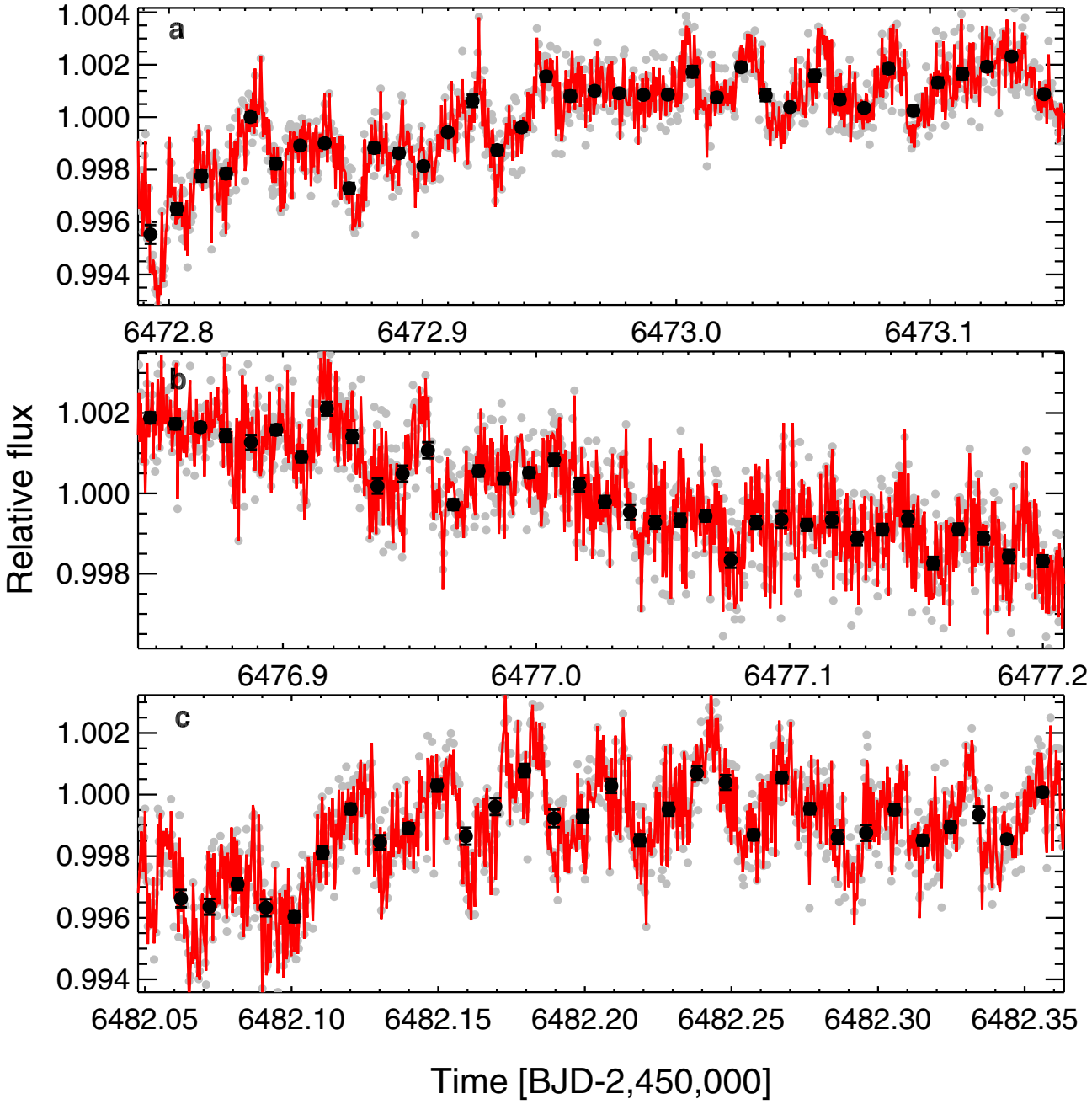
Code availability. The code used to perform the aperture photometry on the Spitzer data sets presented here is publicly available from the IDL Astronomy User's Library at http://idlastro.gsfc.nasa.gov. We have opted not to make the MCMC code available, but the corrected photometry for each data set is available online as Source Data.

31. Ballard, S. *et al.* Kepler-93b: a terrestrial world measured to within 120 km, and a test case for a new *Spitzer* observing mode. *Astrophys. J.* **790,** 12 (2014).
32. Eastman, J., Siverd, R. & Gaudi, B. S. Achieving better than 1 minute accuracy in the heliocentric and barycentric Julian Dates. *Publ. Astron. Soc. Pacif.* **122,** 935–946 (2010).
33. Landsman, W. B. The IDL Astronomy User's Library. In *Astronomical Data Analysis Software and Systems II* Vol. 52 of *ASP Conf. Ser.* (eds Hanisch, R. J. *et al.*) 246–248 (Astronomical Society of the Pacific, 1993).
34. Agol, E. *et al.* The climate of HD 189733b from fourteen transits and eclipses measured by *Spitzer*. *Astrophys. J.* **721,** 1861–1877 (2010).
35. Beerer, I. M. *et al.* Secondary eclipse photometry of wasp-4b with warm spitzer. *Astrophys. J.* **727,** 23 (2011)
36. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **6,** 461–464 (1978).

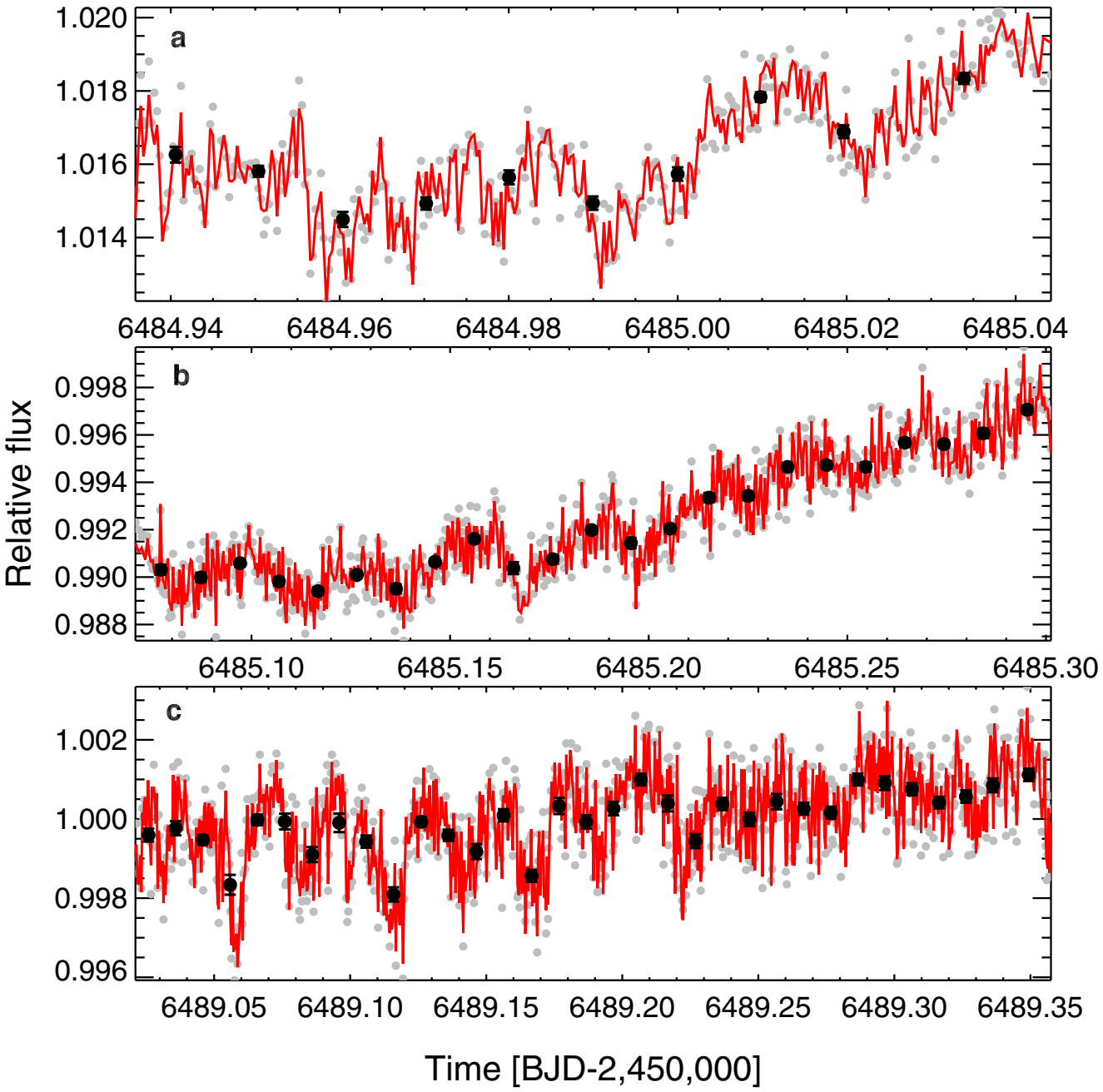
37. Sobolev, V. V. *Light Scattering in Planetary Atmospheres* Vol. 76 of *International Series of Monographs in Natural Philosophy* Ch. 9 (Pergamon Press, 1975) [transl.].
38. Mandel, K. & Agol, E. Analytic light curves for planetary transit searches. *Astrophys. J.* **580,** L171–L175 (2002).
39. Claret, A. & Bloemen, S. Gravity and limb-darkening coefficients for the *Kepler*, CoRoT, *Spitzer*, *uvby*, *UBVRIJHK*, and Sloan photometric systems. *Astron. Astrophys.* **529,** A75 (2011).
40. von Braun, K. *et al.* 55 Cancri: stellar astrophysical parameters, a planet in the habitable zone, and implications for the radius of a transiting super-Earth. *Astrophys. J.* **740,** 49 (2011).
41. Knutson, H. A. *et al.* A map of the day–night contrast of the extrasolar planet HD 189733b. *Nature* **447,** 183–186 (2007).
42. Cowan, N. B. & Agol, E. Inverting phase functions to map exoplanets. *Astrophys. J.* **678,** L129–L132 (2008).
43. Crossfield, I. J. M. ACME stellar spectra. I. Absolutely calibrated, mostly empirical flux densities of 55 Cancri and its transiting planet 55 Cancri e. *Astron. Astrophys.* **545,** A97 (2012).
44. Menou, K. Magnetic scaling laws for the atmospheres of hot giant exoplanets. *Astrophys. J.* **745,** 138 (2012).
45. Owen, J. E. & Wu, Y. Kepler planets: a tale of evaporation. *Astrophys. J.* **775,** 105 (2013).
46. Bolmont, E., Raymond, S. N., Leconte, J., Hersant, F. & Correia, A. C. M. *Mercury-T*: a new code to study tidally evolving multi-planet systems. Applications to Kepler-62. *Astron. Astrophys.* **583,** A116 (2015).



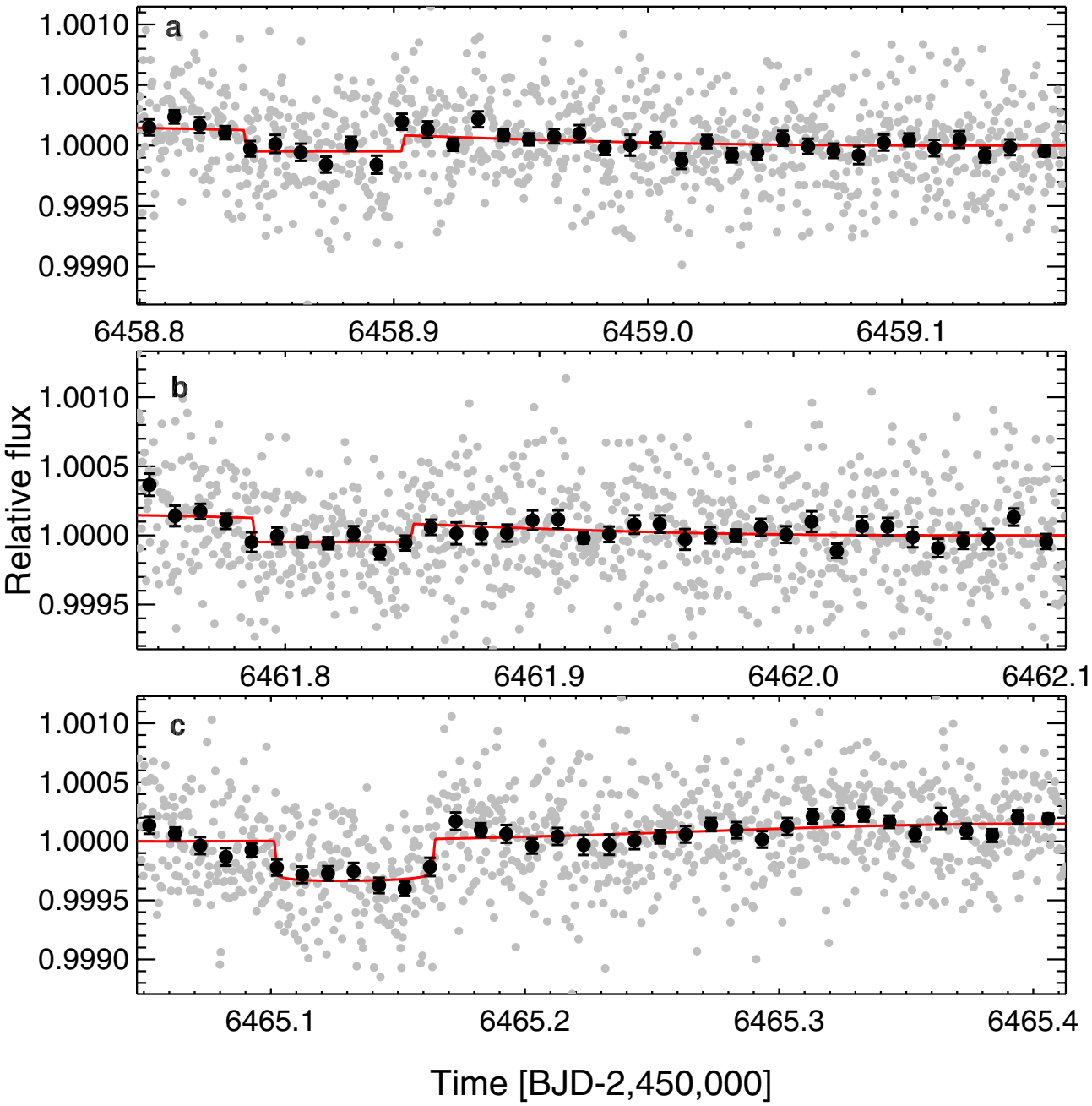
Extended Data Figure 1 | 55 Cancri e raw photometry. a–c, The raw data for time series acquired on 15 June 2013 (**a**), 18 June 2013 (**b**) and 21 June 2013 (**c**). The best-fit instrumental + astrophysical model is superimposed in red. Grey filled circles are data binned per 30 s. Black filled circles are data binned per 15 min. The error bars are the standard deviation of the mean within each time bin. BJD, barycentric Julian date.



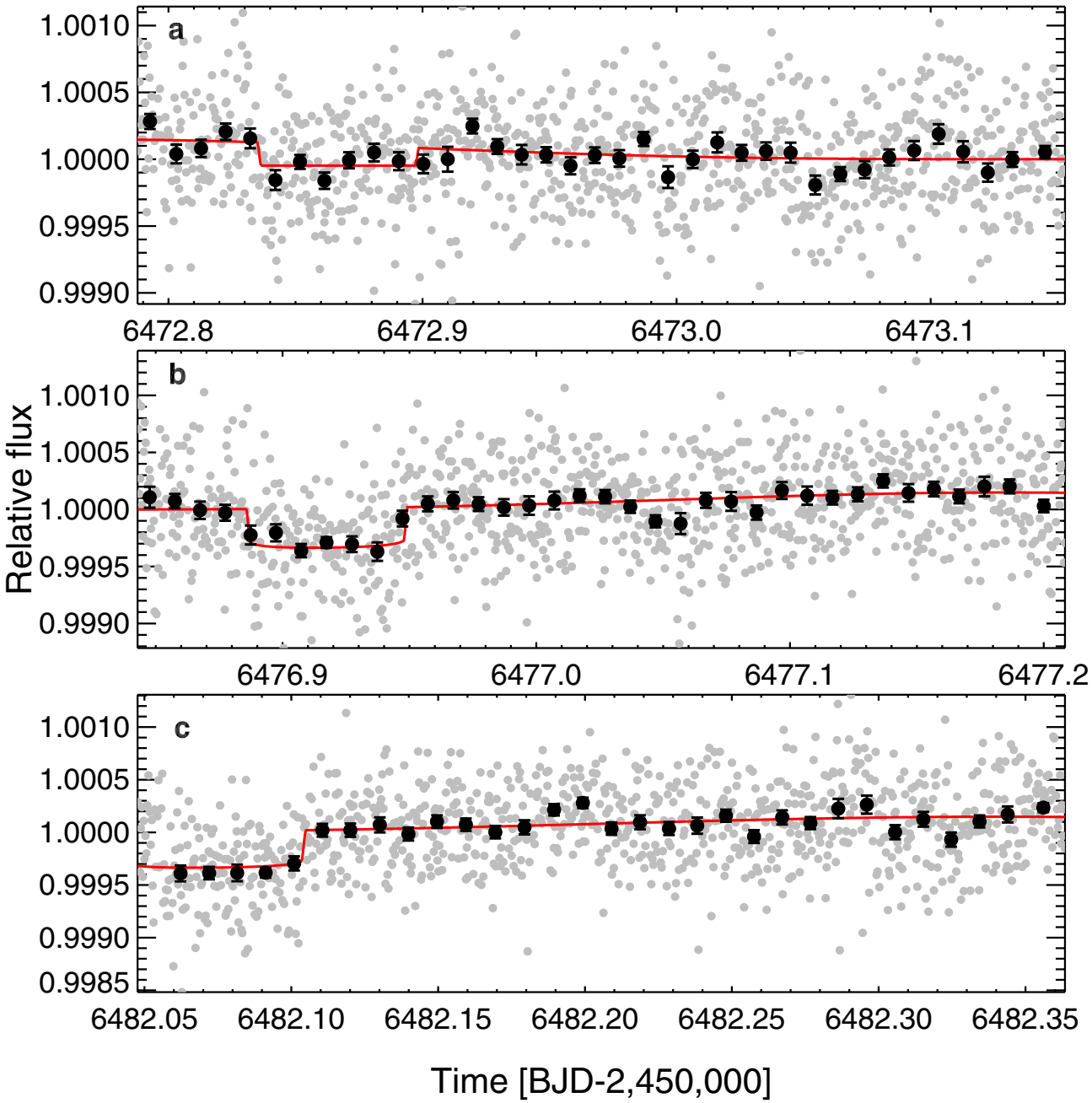
Extended Data Figure 2 | Continuation of Extended Data Fig. 1. a–c, The raw data for time series acquired on 29 June 2013 (**a**), 3 July 2013 (**b**) and 8 July 2013 (**c**).



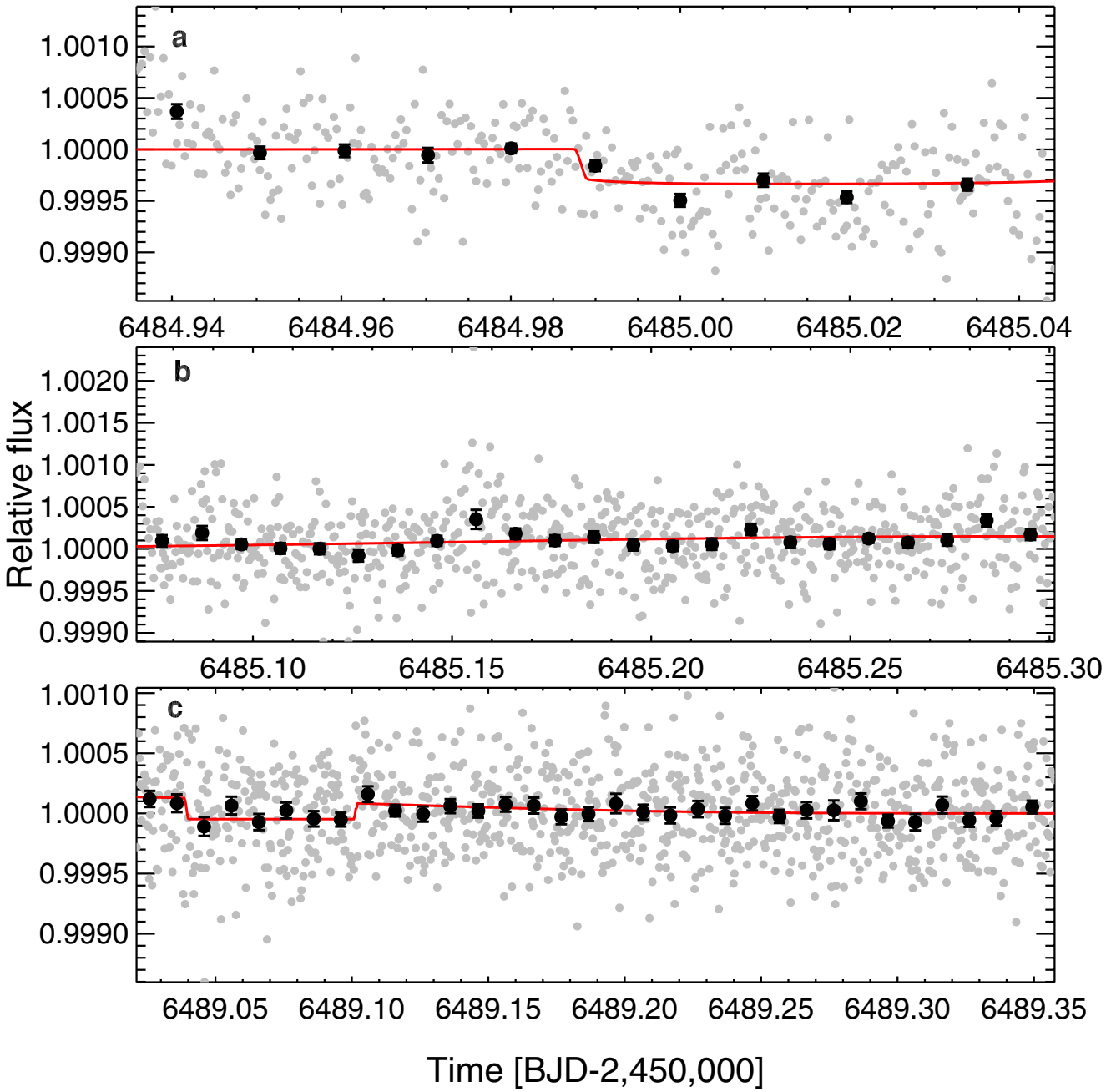
Extended Data Figure 3 | Continuation of Extended Data Fig. 1. a–c, The raw data for time series acquired on 11 July 2013 (**a**, **b**) and 15 July 2013 (**c**).



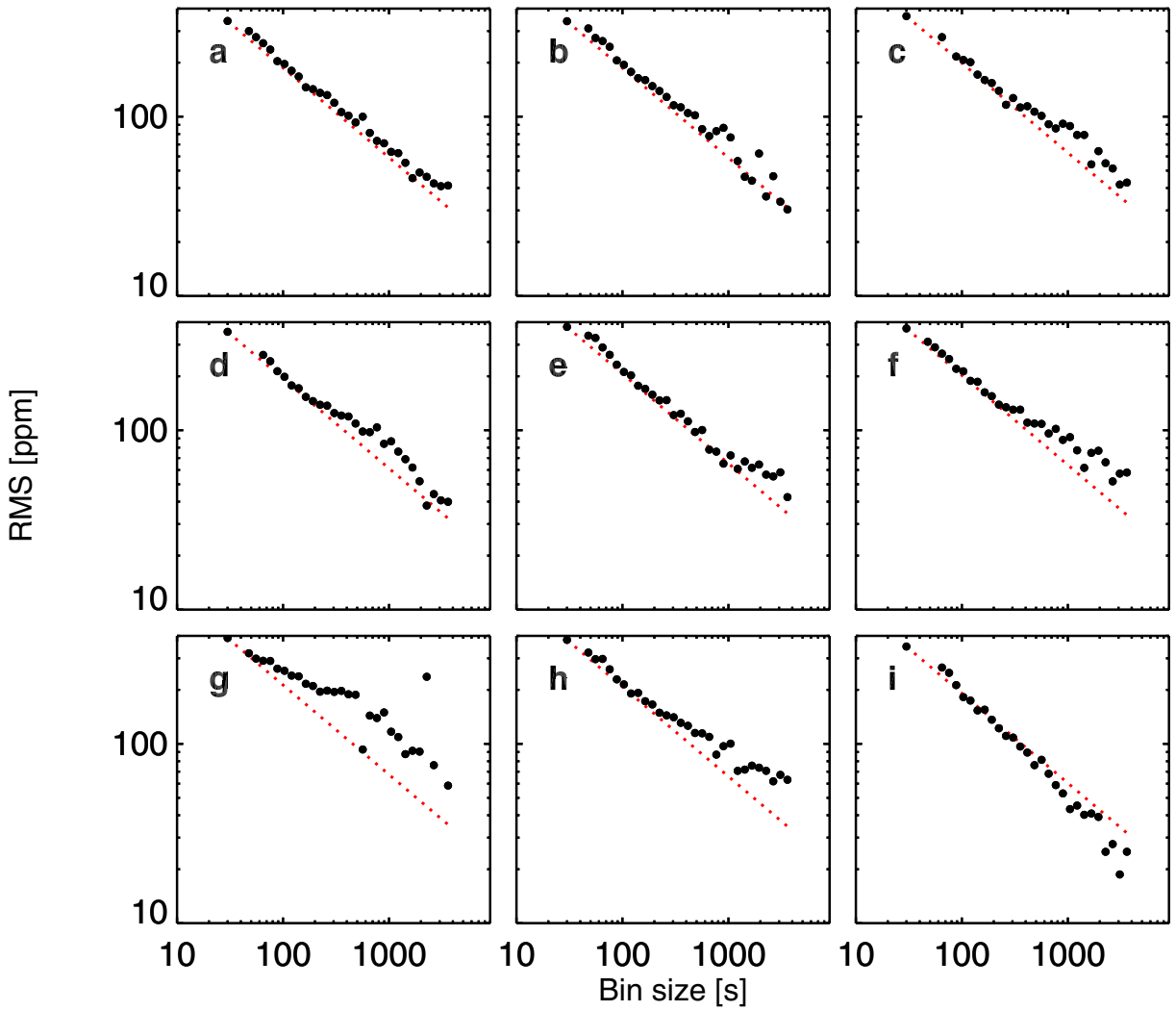
Extended Data Figure 4 | 55 Cancri e corrected photometry. a–c, The detrended data for time series acquired on 15 June 2013 (**a**), 18 June 2013 (**b**) and 21 June 2013 (**c**). The best-fit instrumental + astrophysical model is superimposed in red. Grey filled circles are data binned per 30 s. Black filled circles are data binned per 15 min. The error bars are the standard deviation of the mean within each time bin. BJD, barycentric Julian date.



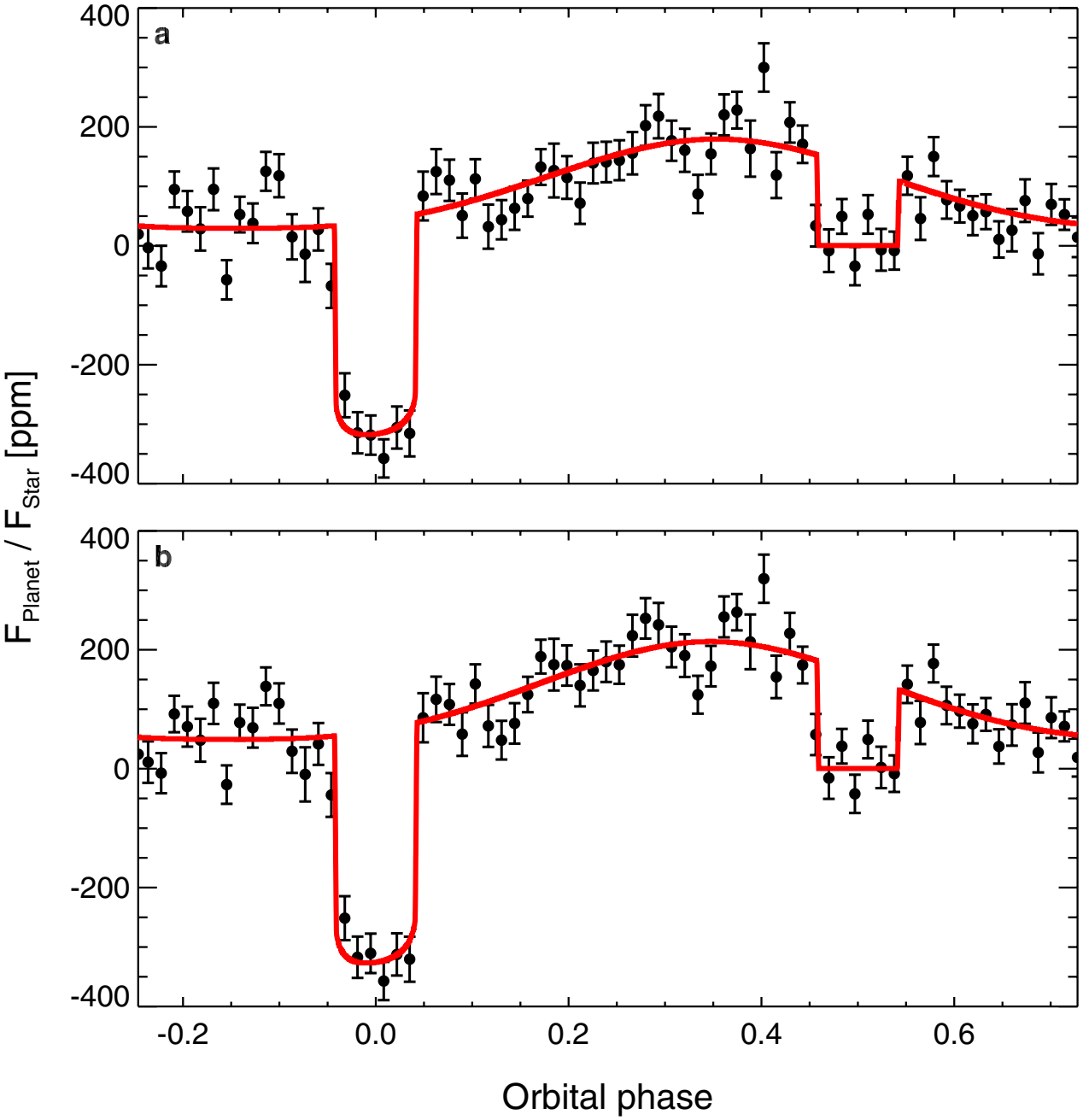
Extended Data Figure 5 | Continuation of Extended Data Fig. 4. a–c, The detrended data for time-series acquired on 29 June 2013 (**a**), 3 July 2013 (**b**) and 8 July 2013 (**c**).



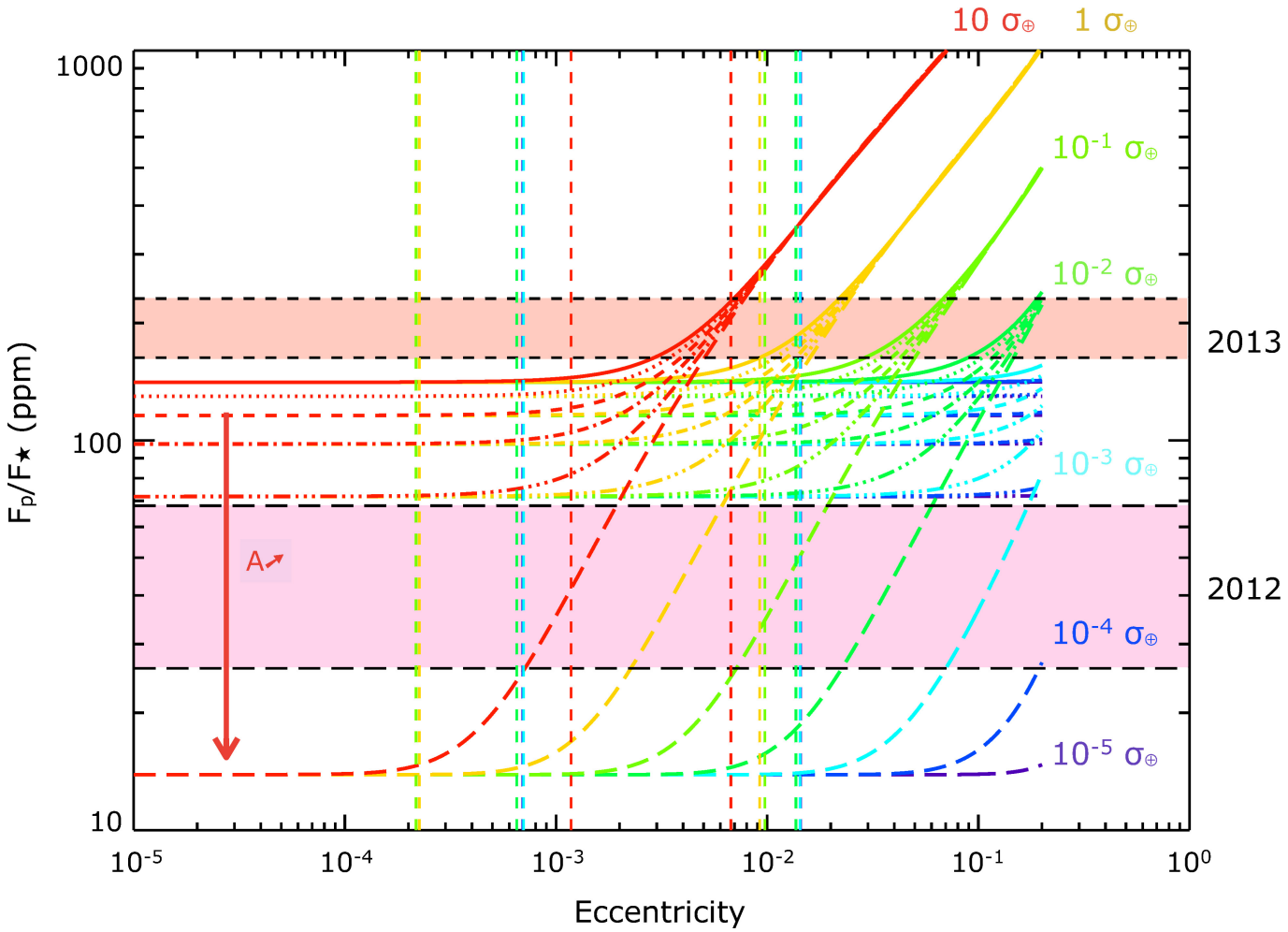
Extended Data Figure 6 | Continuation of Extended Data Fig. 4. a–c, The detrended data for time-series acquired on 11 July 2013 (**a**, **b**) and 15 July 2013 (**c**).



Extended Data Figure 7 | Photometric r.m.s. versus bin size for all data sets. a–i, Black filled circles indicate the photometric residual r.m.s. for different time bins. Each panel corresponds to each individual data set (**a–i**, increasing observing date). The expected decrease in Poisson noise normalized to an individual bin (30 s) precision is shown as a red dotted line.



Extended Data Figure 8 | Polynomial-detrended phase-folded photometry. Photometry for all eight data sets combined and folded on the orbital period of 55 Cancri e. **a**, Fit results using the entire time series as input data. **b**, Fit results obtained by splitting the times series in two. Data in **a** and **b** represent the planet-to-star flux ratio (F_{planet}/F_{star}) variation in phase and are binned per 15 min; the error bars are the standard deviation of the mean within each orbital phase bin. The best-fit model is shown in red. Contrary to Fig. 1, these fits are obtained using polynomial functions of the centroid position and the FWHM of the PRF.



Extended Data Figure 9 | Tidal heating constraints for 55 Cancri e. The planet-to-star flux ratio ($F_p/F_\star$) is shown as a function of the orbital eccentricity for different values of dissipation (relative to the Earth's $\sigma_\oplus$; indicated by the different colours) and albedos ('A', indicated by the different line styles, from 0.0 (solid) to 1.0 (long-dashed)). The pink and orange bands represent the occultation depth values measured in 2012 and 2013 with Spitzer, respectively. Vertical lines indicate the plausible range of the eccentricity of 55 Cancri e as determined from the *N*-body simulations for each dissipation value. The 2012 occultation depth can be matched for high albedos and a high dissipation, while the deeper 2013 occultation depth can be matched for the highest dissipation ($10\sigma_\oplus$) and the whole albedo range.

Extended Data Table 1 | 55 Cancri e Spitzer data set

Date [UT]	Program ID	AOR #	AOR duration [h]	Phase range	Aperture [pix]	Interp. n	RMS/30s [ppm]	β_r
2013-06-15	90208	48070144	8.8	0.40 - 0.89	2.6	64	341	1.17
2013-06-18	90208	48073216	8.8	0.39 - 0.89	3.0	64	340	1.00
2013-06-21	90208	48070656	8.8	0.88 - 0.38	2.8	70	363	1.00
2013-06-29	90208	48073472	8.8	0.39 - 0.89	3.0	58	354	1.16
2013-07-03	90208	48072448	8.8	0.90 - 0.39	3.2	69	376	1.34
2013-07-08	90208	48072704	8.0	0.94 - 0.39	2.6	77	370	1.03
2013-07-11	90208	48072960, p1	2.6	0.88 - 0.03	2.6	22	388	1.77
2013-07-11	90208	48072960, p2	5.5	0.07 - 0.38	2.6	58	389	1.83
2013-07-15	90208	48073728	8.1	0.43 - 0.89	3.4	73	348	1.00

Astronomical Observation Request (AOR) properties for the Spitzer/IRAC 4.5-μm data used here. This table also indicates the planetary orbital phase covered by each AOR as well as the number of interpolation points (n, 'Interp.') relevant to the BLISS algorithm. β_{red}, red-noise contribution to each AOR.

LETTER

doi:10.1038/nature17620

Exploring the quantum speed limit with computer games

Jens Jakob W. H. Sørensen[1], Mads Kock Pedersen[1], Michael Munch[1], Pinja Haikka[1], Jesper Halkjær Jensen[1], Tilo Planke[1], Morten Ginnerup Andreasen[1], Miroslav Gajdacz[1], Klaus Mølmer[1], Andreas Lieberoth[1] & Jacob F. Sherson[1]

Humans routinely solve problems of immense computational complexity by intuitively forming simple, low-dimensional heuristic strategies[1,2]. Citizen science (or crowd sourcing) is a way of exploiting this ability by presenting scientific research problems to non-experts. 'Gamification'—the application of game elements in a non-game context—is an effective tool with which to enable citizen scientists to provide solutions to research problems. The citizen science games Foldit[3], EteRNA[4] and EyeWire[5] have been used successfully to study protein and RNA folding and neuron mapping, but so far gamification has not been applied to problems in quantum physics. Here we report on Quantum Moves, an online platform gamifying optimization problems in quantum physics. We show that human players are able to find solutions to difficult problems associated with the task of quantum computing[6]. Players succeed where purely numerical optimization fails, and analyses of their solutions provide insights into the problem of optimization of a more profound and general nature. Using player strategies, we have thus developed a few-parameter heuristic optimization method that efficiently outperforms the most prominent established numerical methods. The numerical complexity associated with time-optimal solutions increases for shorter process durations. To understand this better, we produced a low-dimensional rendering of the optimization landscape. This rendering reveals why traditional optimization methods fail near the quantum speed limit (that is, the shortest process duration with perfect fidelity)[7–9]. Combined analyses of optimization landscapes and heuristic solution strategies may benefit wider classes of optimization problems in quantum physics and beyond.

Quantum physics could lead to technological advances in the realms of computing[10] and simulations[11]. To ensure functionality, all quantum operations must be executed near perfection, requiring highly optimized operations with fidelities above $F \geq 0.999$ (ref. 12). Given the high dimensionality of quantum optimization problems, one could expect these 'quantum optimal control' problems to be impractically difficult to solve. However, assuming full controllability, quantum optimization problems are benign, since all local maxima are also global maxima[13]. Tailored local optimizers, such as the gradient-based Krotov algorithm, solve these problems[14].

Quantum computing operations must be executed faster than typical decoherence times to ensure functionality. However, there is a shortest process duration with perfect fidelity, denoted the quantum speed limit (QSL)[7], which imposes a fundamental limit on the process duration and hence on quantum computation. With limited duration, the quantum optimization problem loses the favourable properties stated earlier and local optimizers are no longer always guaranteed to converge. In this case the prevalent and hitherto successful method is to use multistart of such local optimization algorithms[8,15,16]. For a time-dependent Hamiltonian the QSL is conventionally computed by assuming that it coincides with the process duration for which the multistarting of local optimization fails[8,16]. We demonstrate that this assumption is not necessarily true.

High-dimensional optimization problems are often solved by humans using simple heuristic strategies. Common examples are visual pattern recognition[17] and catching a flying ball subject to wind and air resistance[1]. Citizen science projects such as Foldit[3], EyeWire[5] and Galaxy Zoo[18] employ these human skills to solve highly complex research problems via gamification. We asked whether citizen science projects can be extended from these puzzle and pattern recognition tasks to dynamic challenges, and whether this approach can be implemented on quantum physics problems. To investigate these questions we created Quantum Moves (at http://www.scienceathome.org; see Supplementary Information), which presents quantum computing operations as games.

The quantum computer architecture we gamify employs neutral atoms trapped in optical lattices[11]. By driving the transition from superfluid to Mott insulator, samples of hundreds of atoms are trapped in optical lattices with unprecedented purity[19,20]. Their regular spacing offers the possibility of creating a scalable quantum computer, which is a challenge for other available architectures[21]. Several proposals for the implementation of quantum computing in this system have been proposed, mainly using long-range gates or contact interactions[11]. Here we investigate an architecture[6] where contact interaction is achieved by moving atoms on top of each other using a so-called optical tweezer[22,23], a tightly focused off-resonant laser beam. Finding the optimal path of the tweezer from one lattice site to another is a difficult problem when the available time is close to the QSL, and it is this transfer problem that is introduced to the players of Quantum Moves through different challenges.

Here we present the results of one challenge called BringHomeWater. The aim of BringHomeWater is to move the optical tweezer into a region where an atom is trapped in a fixed potential well, collect the atom and move it back to a target area as quickly as possible. Fast movement introduces excitations in the state of the atom, which is described by a quantum mechanical wavefunction $\psi(x, t)$ and visualized as a sloshy liquid—see Fig. 1 for the player view of BringHomeWater and Methods for a description of the game interface and player demographics. The excitations must be stabilized before reaching the target area to attain high fidelity of the underlying quantum process.

BringHomeWater belongs to the class of one-dimensional quantum optimal control of individual atoms with one or two control parameters. This class of problems is of great interest because of their relevance for quantum computing and state engineering. These problems have been studied extensively within the past decade (see, for example, refs 6, 9 and 24) and are highly relevant for the experimental platforms in refs 20 and 25. We believe our gamification strategy is extendable to these problems and other physical interactions, including (but not limited to) many-body dynamics in optical lattices[26] and Bose–Einstein condensates[25].

Recall that the standard approach for solving a problem such as BringHomeWater is to use a multistart of tailored local optimization algorithms, such as the Krotov algorithm[8,14,16]. The choice of an initial seed is a central challenge in such complex optimizations. In BringHomeWater

[1]Department of Physics and Astronomy, Aarhus University, Aarhus, Denmark.

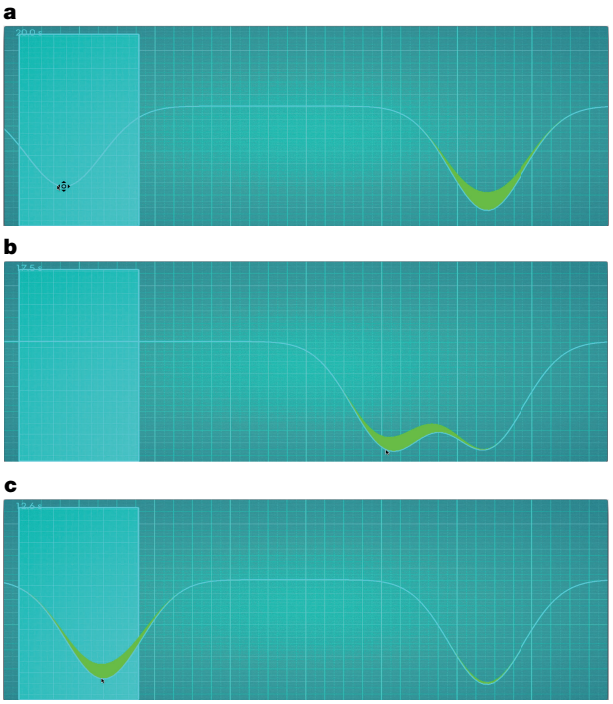


Figure 1 | The BringHomeWater challenge as seen by the player. The atom is represented by the square of its wavefunction, $|\psi(x,t)|^2$, shown as a green liquid. The blue curve represents the potential felt by the atom. The controllable tweezer is initially on the left (**a**) and the atom is trapped in the right static potential. The player controls the optical tweezer by moving a computer cursor, picks up the atom (**b**), and drags it back to the target area (**c**), marked by a cyan rectangle, to collect points.

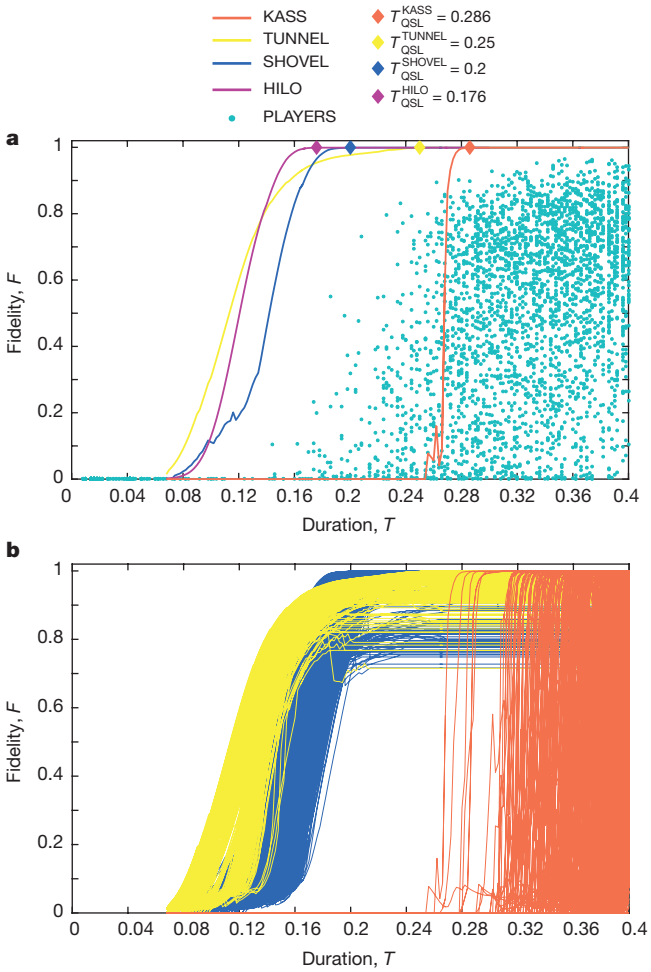


Figure 2 | Fidelities of the transport problem for different solution durations. a, A subset of the solutions found by the players as turquoise dots. The curves show the best solutions found by computer optimizations for each duration. The optimizations shown are KASS (red), shovelling (blue), tunnelling (yellow), and HILO (purple). The diamonds mark the shortest duration with optimal fidelity ($F \geq 0.999$) for each optimization method. **b**, The sweeps from seeds that are generated by players (yellow and blue) and computers (red). Player solutions divide into shovelling (blue) and tunnelling (yellow) clans.

the high-dimensional optimization space can only be searched sparsely, and we used multiple methods for creating initial seeds for the Krotov algorithm. The most successful method, KASS, employs linear combinations of sinusoidal functions as seed trajectories (see Methods). Using the Krotov algorithm on these seeds, high-fidelity solutions are readily found for process durations longer than $T = 0.40$ (units defined in Methods). Iteratively, solutions with shorter durations are found by contracting the solutions with a slightly longer duration and using them as seeds for the Krotov algorithm. This procedure is called a sweep, and it traces out entire families of solutions. The best results of KASS are displayed in red in Fig. 2a. This method locates a numerical estimate of the QSL at $T_{QSL}^{num} = 0.29$ after approximately 7.4×10^8 trials. This KASS optimization formed our most successful bare computer optimization method, outperforming also the acclaimed CRAB algorithm[27].

KASS and CRAB fit the prevalent paradigm of multistarting of local optimizers. However, global optimizers are rarely used in quantum optimal control. To investigate the BringHomeWater problem using a global optimizer we chose the so-called differential evolution algorithm because of its demonstrated success in quantum problems[28]. For duration $T = 0.40$, differential evolution performed worse than the Krotov optimization with identical computational resources. The poor performance of differential evolution can be attributed to the very high dimensionality of the optimization space, and to the scarceness of good solutions found therein.

Players were trained in a series of introductory levels before reaching the scientific challenges. Training levels equip the players with a range of skills; successful solutions of the scientific challenges require a holistic combination of these skills. In total, all Quantum Moves games have been played about 500,000 times by roughly 10,000 players. BringHomeWater is the most-played scientific level (approximately 12,000 plays by 300 players) and is therefore analysed in detail here. Player results span many different fidelities and process durations—see the dots in Fig. 2a. Remarkably, players trace out a region in the vicinity of T_{QSL}^{num} that is very similar to that of the best numerical results obtained, despite the fact that the numerical optimization used roughly 100,000 times more trials than the players. For short durations, players find even better solutions than the numerical optimization, albeit with imperfect fidelities.

This result inspired us to introduce a powerful hybrid Computer–Human Optimization (CHOP) scheme, in which we use the players' intuitive solutions as seeds for the local numerical optimization. CHOP was applied to the top 70% of the player solutions with durations shorter than $T = 0.40$. The results of these player-seeded optimization sweeps are shown in Fig. 2b. It turns out that the CHOP solutions bunch in groups, which we denote 'clans' and discuss in detail later. In Fig. 2a we illustrate the best solution families of the two dominant clans (blue and yellow curves). The QSL found by CHOP is $T_{QSL}^{num} = 0.20$, a vast improvement on the value obtained from the bare computer optimization using multistarted local optimization. This result clearly demonstrates that BringHomeWater is a quantum control problem without the benign properties discussed previously and questions the common assumption that the QSL can be found numerically as the duration for which local optimization fails[16].

To understand better how the players and CHOP can give better solutions than KASS, we examine the clustering of the CHOP solutions.

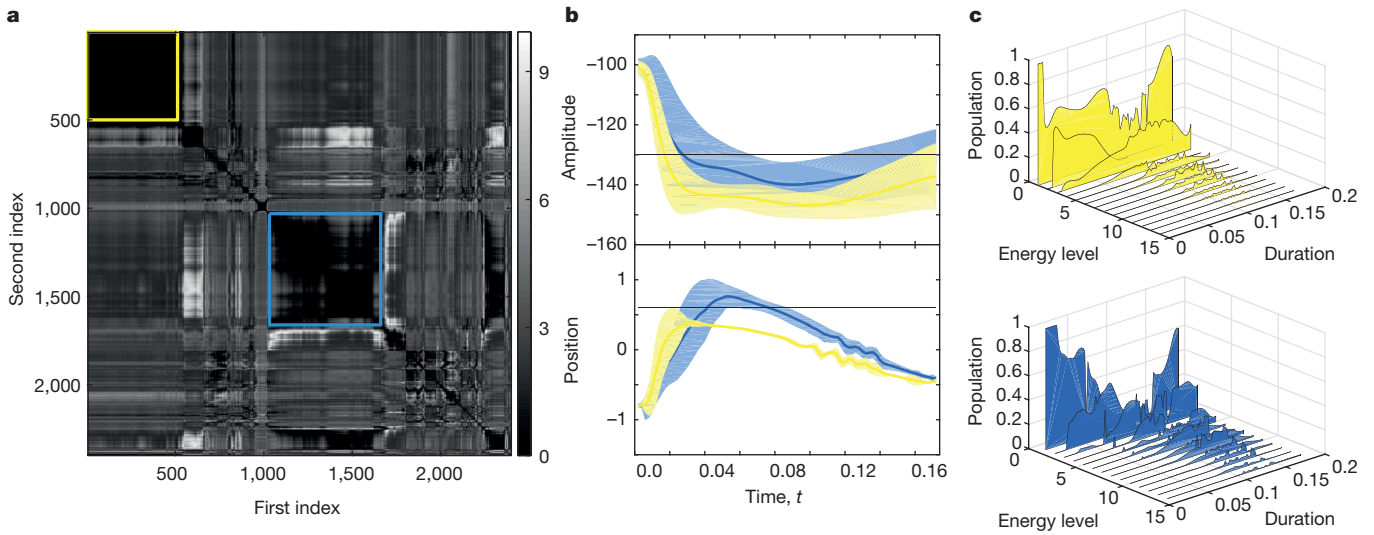


Figure 3 | Shovelling and tunnelling clans. The shovelling (tunnelling) clan is marked blue (yellow) throughout. **a**, Distance map showing the distances between CHOP solutions as defined by equation (1). Boundaries of the clans are marked with coloured squares. **b**, Average trajectories followed by the clans are shown as thick lines, and single standard deviations thereof as translucent areas. Trajectories are divided into the tweezer amplitude (top) and tweezer position (bottom). **c**, The wavefunction fraction (population) in the different instantaneous energy eigenstates.

We introduce a measure describing the distance between two solutions for a particular process duration T:

$$D_{j,k} = \frac{1}{T}\int_0^T \left\langle f_{jk} \middle| f_{jk} \right\rangle \mathrm{d}t \qquad (1)$$

where $|f_{jk}\rangle = |\psi_j(x,t)\rangle - |\psi_k(x,t)\rangle$ is the difference between two wavefunctions evaluated along the path. Figure 3a shows a distance map of the CHOP solutions. Two dominant clans stand out as distinct regions in the distance map; see Methods for details on the identification of clans. The solutions of the BringHomeWater transport problem corresponding to these two clans are shown in Fig. 3b. Using the first solution, marked in yellow in Fig. 3b, the atom is collected by 'tunnelling' the wavefunction into a tweezer potential placed on the left-hand side of the static potential. In the second class of 'shovelling' solutions, marked in blue in Fig. 3b, the tweezer is moved past the position of the atom so that the overlap of the tweezer and the static potential forms a strong potential gradient accelerating the atom towards the target. Fast non-adiabatic solutions must spread the wavefunction into different energy eigenstates. In Fig. 3c we see that the solutions in both clans populate many different eigenstates. The tweezer must then be shaken periodically when approaching the target area to bring the atom back to the desired ground state—see Methods for more details. *A priori*, it was not obvious that two strategies, corresponding to distinctly different physical phenomena, should exist for this kind of a problem. It is also worth stressing that players explore both solutions despite having no or little prior knowledge of quantum mechanical phenomena such as tunnelling.

The difficulty of a particular optimization problem can often be assessed with knowledge of the topology or ruggedness of the so-called optimization or fitness landscape. This has been established for problems in quantum optimal control[29] and in other fields[30]. In this landscape the quality of the solution for each set of the control parameters is represented as the height. If many global optima are distributed across the landscape, local gradient-based search is often sufficient to find the highest peak. On the other hand, if the landscape is very rugged and contains many local maxima, local methods will in general fail. To understand our problem using this terminology we performed a dimensional reduction of the data from the player-seeded solutions and the bare numerical optimization. We assign points to individual solutions in a two-dimensional space such that the Euclidean distance between two points approximates the Manhattan type distance between two solutions in the full high-dimensional optimization space (see Methods). The height of the landscape quantifies the fidelity of a solution.

For process durations $T = 0.40$ and $T \approx 0.17$, above and below $T_{\mathrm{QSL}}^{\mathrm{num}}$, the landscapes are visualized in Fig. 4a and b, respectively. Green areas in the landscape mark the player-generated CHOP solutions. For the longer duration (Fig. 4a), global maxima indeed spread across the optimization landscape, explaining the success of Krotov-based methods. For the shorter duration (Fig. 4b) all high-fidelity solutions lie in the green CHOP region. This explains the failure of the Krotov-based methods and difficulty in locating the true T_{QSL}, since the global optima are no longer spread across the landscape. Thus Fig. 4a and b demonstrates the dramatic change in the landscape as the duration is decreased. We emphasize that the CHOP region is tiny in the high-dimensional landscape and therefore easily missed even by elaborate seeding strategies. CHOP outperforms the numerical optimization because players are able to heuristically identify the regions of high fidelity in the high-dimensional optimization landscape, thereby finding the best seeds for Krotov-type local algorithms.

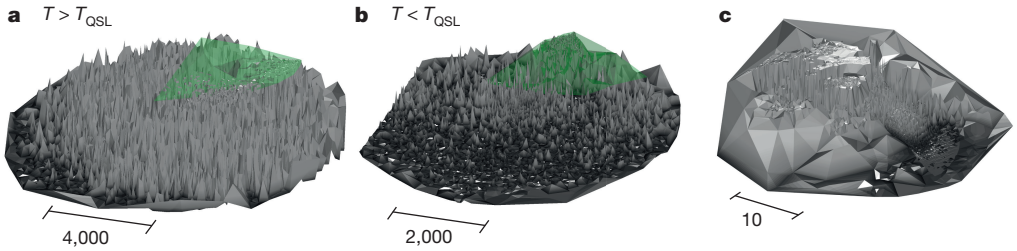


Figure 4 | Optimization landscapes. a, **b**, The two-dimensional rendering of the high-dimensional optimization landscape for process durations $T = 0.40$ and $T = 0.17$, respectively. Green areas mark the space probed by CHOP solutions. **c**, The low-dimensional HILO landscape.

One long-term goal of our work is to circumvent the need for gamification by learning how players form their successful low-dimensional heuristic strategies, and to incorporate this into autonomous optimization algorithms. As a first step in this direction, we introduce here a Heuristically Initialized Local Optimizer (HILO) algorithm. HILO parameterizes the player solutions in a low-dimensional subspace, while retaining the main features of good seeds. A local search algorithm can then move beyond this subspace to find optimal solutions. More specifically, inspired by the player solutions, we constructed a three-dimensional parameterization consisting of moving the tweezer (1) right, (2) slowly left, and (3) quickly left (see Methods for details). Paths from this three-dimensional space for $T=0.15$ were used as seeds for the Krotov algorithm and iteratively applied to shorter and longer durations with a sweep (purple curve in Fig. 2a). Seeds were taken from the low-dimensional subspace using a simple direct search. HILO finds the lowest QSL at $T_{\mathrm{QSL}}^{\mathrm{num}}=0.176$, outperforming even the best CHOP strategies. The solutions from HILO are shown as a landscape in Fig. 4c. Initialized in a low-dimensional seed space, HILO efficiently explores a smaller but more optimal volume of the global optimization space. Only parameterizations that accurately capture the nature of efficient solution strategies at short durations will lead to efficient optimization. The intuition gained from the players was pivotal, as the parameterization used in HILO emerged from the CHOP solutions and the physical interpretation of the player-based solution strategies. This makes CHOP a crucial precursor to HILO. We stress that any optimal control effort requires a substantial amount of optimizations of different seeds to find good solutions. After this initial work, HILO could be readily implemented by first applying our clustering methods to identify potential clans. If successful, analysis of the clans would allow for a formulation of the low-dimensional parameterization used in HILO.

Finally, we discuss another central component of quantum optimal control theory, namely the dependence of the fidelity on the process duration around the QSL. This has previously been associated with a universal-type $\sin^2\left(\frac{\pi}{2}T/T_{\mathrm{QSL}}\right)$ behaviour[8]. Here we do not observe this behaviour (see Extended Data Fig. 1a and Methods for details). We attribute this to variations in the so-called direct Hilbert velocity[9]. The variations in our problem may arise from the sequential nature of the good strategies, such as the three steps in the HILO parameterization, where each sequence has a different scaling of fidelity with time.

Using the gamified interface in Quantum Moves, players with little or no training in quantum physics not only provided high-quality solutions, but also enabled the extraction of the underlying physical strategies. This success encourages the pursuit of other quantum research games, as well as dynamic games in other fields. Our future work will focus on extending the gamification strategy and also the more general classification methodology to new types of control problems such as those presented in refs 24–26. For all these different physical interactions we expect duration-constrained problems to exhibit complex optimization landscapes. Here players are expected to provide a global overview beyond typical local optimization. Another interesting topic of future research will be the exploration of how the players form strategies. Analyses of the player data will enable us to identify important features of quantum optimal control problems and allow an efficient dimensionality reduction. This will form the first steps in the efficient training of modern machine learning algorithms. As a first step in this direction, we have collaborated with cognitive researchers to create a new game called Quantum Minds, also available at www.scienceathome.org.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 9 July 2015; accepted 12 February 2016.

1. McLeod, P. & Dienes, Z. Do fielders know where to go to catch the ball or only how to get there? *J. Exp. Psychol. Hum. Percept. Perform.* **22,** 531–543 (1996).
2. Gigerenzer, G. & Todd, P. in *Simple Heuristics That Make Us Smart* (eds Gigerenzer, G., Todd, P. & The ABC Research Group) 3–34 (Oxford Univ. Press, 1999).
3. Cooper, S. *et al.* Predicting protein structures with a multiplayer online game. *Nature* **466,** 756–760 (2010).
4. Lee, J. *et al.* RNA design rules from a massive open laboratory. *Proc. Natl Acad. Sci. USA* **111,** 2122–2127 (2014).
5. Kim, J. S. *et al.* Space-time wiring specificity supports direction selectivity in the retina. *Nature* **509,** 331–336 (2014).
6. Weitenberg, C., Kuhr, S., Mølmer, K. & Sherson, J. Quantum computation architecture using optical tweezers. *Phys. Rev. A* **84,** 032322 (2011).
7. Mandelstam, L. & Tamm, I. The uncertainty relation between energy and time in non-relativistic quantum mechanics. *J. Phys.* **9,** 249–254 (1945).
8. Caneva, T., Calarco, T., Fazio, R., Santoro, G. E. & Montangero, S. Speeding up critical system dynamics through optimized evolution. *Phys. Rev. A* **84,** 012312 (2011).
9. Gajdacz, M., Das, K. K., Arlt, J., Sherson, J. F. & Opatrný, T. Time limited optimal dynamics beyond the quantum speed limit. *Phys. Rev. A* **92,** 062106 (2015).
10. Monroe, C. Quantum information processing with atoms and photons. *Nature* **416,** 238–246 (2002).
11. Lewenstein, M. *et al.* Ultracold atomic gases in optical lattices: mimicking condensed matter physics and beyond. *Adv. Phys.* **56,** 243–379 (2007).
12. Devitt, S. J., Munro, W. J. & Nemoto, K. Quantum error correction for beginners. *Rep. Prog. Phys.* **76,** 076001 (2013).
13. Rabitz, H., Hsieh, M. M. & Rosenthal, C. M. Quantum optimally controlled transition landscapes. *Science* **303,** 1998–2001 (2004).
14. Sklarz, S. & Tannor, D. Loading a Bose-Einstein condensate onto an optical lattice: an application of optimal control theory to the nonlinear Schrödinger equation. *Phys. Rev. A* **66,** 053619 (2002).
15. Ugray, Z. *et al.* Scatter search and local NLP solvers: a multistart framework for global optimization. *Inf. J. Comp.* **19,** 328–340 (2007).
16. Caneva, T. *et al.* Optimal control at the quantum speed limit. *Phys. Rev. Lett.* **103,** 240501 (2009).
17. Bilalić, M., Langner, R., Erb, M. & Grodd, W. Mechanisms and neural basis of object and pattern recognition: a study with chess experts. *J. Exp. Psychol. Gen.* **139,** 728–742 (2010).
18. Lintott, C. *et al.* Galaxy Zoo 1: data release of morphological classifications for nearly 900,000 galaxies. *Mon. Not. R. Astron. Soc.* **410,** 166–178 (2011).
19. Bakr, W. S. *et al.* Probing the superfluid-to-Mott insulator transition at the single-atom level. *Science* **329,** 547–550 (2010).
20. Sherson, J. F. *et al.* Single-atom-resolved fluorescence imaging of an atomic Mott insulator. *Nature* **467,** 68–72 (2010).
21. Ladd, T. D. *et al.* Quantum computers. *Nature* **464,** 45–53 (2010).
22. Weitenberg, C. *et al.* Single-spin addressing in an atomic Mott insulator. *Nature* **471,** 319–324 (2011).
23. Kaufman, A. *et al.* Entangling two transportable neutral atoms via local spin exchange. *Nature* **527,** 208–211 (2015).
24. De Chiara, G. *et al.* Optimal control of atom transport for quantum gates in optical lattices. *Phys. Rev. A* **77,** 052333 (2008).
25. Jäger, G., Reich, D. M., Goerz, M. H., Koch, C. P. & Hohenester, U. Optimal quantum control of Bose–Einstein condensates in magnetic microtraps: comparison of gradient-ascent-pulse-engineering and Krotov optimization schemes. *Phys. Rev. A* **90,** 033628 (2014).
26. Doria, P., Calarco, T. & Montangero, S. Optimal control technique for many-body quantum dynamics. *Phys. Rev. Lett.* **106,** 190501 (2011).
27. Caneva, T., Calarco, T. & Montangero, S. Chopped random-basis quantum optimization. *Phys. Rev. A* **84,** 022326 (2011).
28. Zahedinejad, E., Schirmer, S. & Sanders, B. C. Evolutionary algorithms for hard quantum control. *Phys. Rev. A* **90,** 032310 (2014).
29. Roslund, J. & Rabitz, H. Experimental quantum control landscapes: inherent monotonicity and artificial structure. *Phys. Rev. A* **80,** 013408 (2009).
30. Vuculescu, O. & Bergenholtz, C. How to solve problems with crowds: a computer-based simulation model. *Creativity Innov. Manage.* **23,** 121–136 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the Quantum Moves players, without whom this work would have been impossible. We thank J. Rafner for graphical support and J. Jarecki, O. Vuculescu and C. Bergenholtz for discussions. This work was supported by the European Research Council, the Lundbeck Foundation, the Aarhus University Research Foundation, the Templeton Foundation, the Danish Council for Independent Research, the Villum Foundation and the Carlsberg Foundation.

Author Contributions All authors contributed to the construction of the online game platform and the effort to enlist users. J.J.W.H.S., M.K.P., T.P., M.G.A., M.G., K.M. and J.F.S. participated in the numerical analysis of the player and computer results. All authors contributed to the writing of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.F.S. (sherson@phys.au.dk).

METHODS

Quantum Moves game interface. In the online computer game Quantum Moves, the potential created by the optical tweezer is represented along the horizontal (position) and vertical (energy) axes (see Fig. 1a). The potential is updated as the player uses the computer cursor to control the position and intensity of the optical tweezer. The sloshy liquid seen by the player on top of the tweezer potential is the probability density of the atom position. The probability density ($|\psi(x,t)|^2$) is defined by the quantum mechanical wavefunction $\psi(x,t)$. The wavefunction is updated using the time-dependent Schrödinger equation, according to the player-controlled potential. Thus, the time evolution is both observed and affected in real time by the player. The time is scaled approximately by a factor of 3×10^4, so that atomic evolution at microsecond timescale is experienced over a typical 10-s execution time of the individual game.

The aim of the games is to find the fastest possible path that transfers the atom into the (stationary) ground state with unit fidelity, $F=1$, in the specified target region. When the tweezer is moved quickly, the atomic probability distribution begins to visibly slosh in a fluid-like fashion and the players may readily adopt strategies to prevent or control excitation (sloshing) of the wave. To encourage players to search for the fast solutions necessary for realistic quantum computations, and to probe the region around the QSL, we introduce a time penalty in the game structure. For a thorough discussion on the Quantum Moves player demographics, including their scientific backgrounds, we refer the reader to ref. 31 and for a similar analysis of Zooniverse.org players see ref. 32.

KASS. We tried different seeding and optimization strategies, and the most successful was a Krotov algorithm using sinusoidal seed functions and sweeps over the total duration (KASS). KASS applies a sweep to an initial random seed $u_i(t)$ where $u_1(t)=\mathcal{A}(t)$ is the tweezer amplitude and $u_2(t)=x(t)$ is the tweezer position. The initial random seed is:

$$u_i(k\delta t)=w_i(k\delta t)+\sum_{n=0}^{N-1}X[n]\sin\left(\frac{n\pi k}{N}\right) \qquad (2)$$

where N is the number of time slices in the path and $\delta t=0.002$ is the time discretization. w_i is the motion at constant speed to the initial position of the atom and back again for $\mathcal{A}(t)=-100$. The amplitudes $X[n]$ were selected with a random sign and a norm decreasing as a function of the frequency. We generated roughly 2,400 seeds using the random summation series and apply the Krotov algorithm to these. The next step in KASS is to apply a sweep to the random seeds $u_i(t)$:

(1) The optimized solution for the tweezer depth and position is linearly contracted in time by δt, $\mathcal{A}(t)$, $x(t)\rightarrow\mathcal{A}(at)$, $x(at)$ for $0<a<1$.
(2) The Krotov algorithm is applied using the contracted solution as seed.
(3) Repeat steps (1) and (2) until a minimum duration of $T=0.07$ is reached.

Each of the 2,400 optimizations took 6 h of calculation.

Schrödinger equation in dimensionless units. The time-dependent Schrödinger equation solved in Quantum Moves has the dimensionless form:

$$-\frac{1}{2}\frac{\mathrm{d}^2\psi}{\mathrm{d}x^2}+V(x,t)\psi=i\frac{\mathrm{d}\psi}{\mathrm{d}t}$$

where the position x is measured in units of $l_{\mathrm{unit}}=\lambda/2=532\,\mathrm{nm}$. This is a typical period of optical lattice potentials used in experiments with ^{87}Rb atoms[20]. The potential energy $V(x,t)$ is given in units of $E_{\mathrm{unit}}=1.70\,\mathrm{peV}$, about a fifth of the so-called recoil energy in the lattice potential. Time is given in units of $t_{\mathrm{unit}}=\hbar/E_{\mathrm{unit}}=0.39\,\mathrm{ms}$. In these dimensionless units, our optical tweezer is parameterized as:

$$V_{\mathrm{tweezer}}=\mathcal{A}\exp\left(-\frac{2.0(x-x_0)^2}{w_0^2}\right)$$

where x_0 is the position and w_0^2 is the waist. In the BringHomeWater challenge we have chosen $\mathcal{A}=130$ for the depth of the static tweezer potential and $w_0=0.25$ for the width of both tweezer potentials. The time-dependent Schrödinger equation is solved using the split-step method.

Identification of solution clans. The distance between solutions was calculated using the integral of equation (1) for $T\approx0.17$, the duration for which the two best CHOP families in Fig. 2b intersect. The solutions were then sorted by picking one state at random, and choosing the next state on the list to be the closest one, in the sense of equation (1). This process is iterated with the $(n+1)$th solution chosen as the one closest to the nth solution. The reachability plot of Extended Data Fig. 2 shows the distances between the $(n+1)$th and the nth solutions. Valleys in the reachability plot clearly identify blocks of closely spaced solutions, constituting our solution clans. Clans (valleys) were selected by setting a minimum clan size to 200 and an upper threshold 0.05 for the distance between consecutive solutions in a clan. These clans are the clans marked in yellow and blue in Fig. 3a and Extended Data Fig. 1. As the initial solution is chosen at random, this gives N distinct ways of sorting the solutions. We found that the large clans were essentially independent of the choice of the initial solution used for the sorting.

Physical interpretation of solution strategies. In the tunnelling clan, marked yellow in Figs 2 and 3 and Extended Data Fig. 3, the atom is collected by tunnelling it into the tweezer potential, which is placed on the left-hand side of the static potential. As illustrated in Fig. 3b, around the time $t=0.06$, all of the ~500 yellow solutions move to a very particular location in space, which we interpret as the position maximizing the tunnelling rate. Instead, the depth of the tweezer potential at this position is only weakly correlated with the final fidelity of the solution (Fig. 3b), permitting large variations in its values. This strategy fails for short durations ($T\leq0.22$) because there is no longer time for the atom to tunnel completely between the potential wells. Although seemingly simple, this strategy is at variance with the intuition obtained from similar tunnelling-based problems[33–35], which require careful resonant matching of initial- and final-state energy levels. In contrast, the CHOP solution utilizes a deep potential for the transport tweezer. Combined with precise motional control of the tweezer, this loads the atom directly into a state with high energy spread (see Fig. 3c).

In the shovelling clan, marked blue in Figs 2 and 3, the tweezer is moved past the position of the atom. Therefore the overlap of the tweezer and the static potential forms a strong potential gradient that accelerates the atom towards the target (see Fig. 3b). This strategy transfers the atom into a superposition of many different instantaneous energy levels with a large energy spread ΔE (see Fig. 3c), allowing fast motion in Hilbert space towards the target state. As illustrated in the inset of Extended Data Fig. 1a, this shovelling strategy stays optimal for shorter durations than the tunnelling strategy. Below $T=0.20$, however, it cannot be executed owing to a physical bound imposed on the speed of the tweezer, and the yellow tunnelling clan becomes superior.

Mapping the high-dimensional landscape to two dimensions. The visualization of the control landscape maps all solutions to a two-dimensional surface. The mapping aims to represent distances $D_{j,l}$ between any two solutions. This distance is a Manhattan type distance:

$$D_{j,l}=\sum_i\int_0^T|u_i^j-u_i^l|\mathrm{d}t$$

where u_i^j and u_i^l is a pair of solutions and i=1, 2 denotes the rescaled tweezer position and amplitude to intervals of unit length. The solutions are mapped to the two-dimensional landscape by the construction rules:

(1) A randomly chosen solution is placed at the origin $(x_1,y_1)=(0,0)$.
(2) The two solutions closest to the initial solution are given Euclidean coordinates (x_i,y_i) such that Euclidean distances between them match the Manhattan type distances $D_{1,2}$, $D_{1,3}$, and $D_{2,3}$. The points define a triangle with an arbitrary orientation.
(3) The solution with the smallest Manhattan distance to the previous ones is given coordinates (x_k,y_k) such that the two-dimensional distances are as close as possible to the values $D_{i,k}$ $i=1,2,\ldots,k-1$. This is done with the Nelder–Mead algorithm using the cost function:

$$S_k=\sum_{j=1}^{k-1}\left|D_{j,k}^{\mathrm{E}}-D_{j,k}\right| \qquad (3)$$

where $D_{j,k}^{\mathrm{E}}$ denotes the Euclidean distance between the coordinates of two solutions in a two-dimensional landscape.

A three-dimensional figure is obtained by plotting (x_i,y_i,F_i) where F_i is the fidelity of the ith solution. This procedure gives a qualitative impression of the optimization landscape. Landscapes are constructed from player seeds and random computer seeds; optimized solutions at the durations $T\approx0.17$ and $T=0.40$ are shown in Fig. 4. The multitude of spikes appears because gradient-based optimization leads to nearly vertical lines in the landscape. This highlights the failure of local optimization algorithms to explore extended parts of the global landscape. A similar landscape for HILO is also shown in Fig. 4.

HILO. HILO, or Heuristically Initialized Local Optimizer, applies the Krotov algorithm to a seed found using the same heuristic as the best CHOP solutions. The motion of the tweezer in a tunnelling (yellow) and shovelling (blue) CHOP solution can be reasonably approximated by three movements at constant horizontal speed and rate of change of the potential depth (see Extended Data Fig. 3), which connects three points $P_i=(x_i,A_i,t_i)$, for $i=1,2,3$. These three lines correspond to the first quick movement to a point P_1, from P_1 a slow backwards motion to P_2 and a final move to P_3 inside the target area. This defines an optimization problem with dimension $D=9$. P_3 was set equal to the last point in w_i. We noticed that the duration of the quick movement (t_1) should be as small as possible, reducing the seed space dimension to $D=5$. This dimension could be further reduced to $D=3$

by only changing the position of the tweezer using straight lines and letting the amplitude decay exponentially to $\mathcal{A}=-150$ after t_1. Optimizations showed that $\mathcal{A}=-150$ was the optimal value. To reach the desired state with high fidelity by shaking the tweezer is crucial. This shaking is deliberately not included in the parameterization, since the Krotov algorithm is excellent at introducing this type of motion by itself. We generated seeds for the Krotov algorithm from this space using a direct search.

The direct Hilbert velocity. In numerous studies the fidelity below the QSL has been found to follow a universal $\sin^2\left(\frac{\pi}{2}T/T_{\mathrm{QSL}}\right)$ behaviour[8]. This section briefly summarizes how an expression for the fidelity can be obtained using the concept of direct Hilbert velocity Q (ref. 9). State $|\psi(0)\rangle$ is evolved in time under a controlled Hamiltonian $\hat{\mathcal{H}}(t)$ towards a target state $|\chi(T)\rangle$ at time T. The Krotov algorithm exploits the fact that the target state can be propagated backwards in time using the (adjoint) time evolution operator. Using $|\chi(T)\rangle$, the process fidelity can be evaluated at any instant of time:

$$F=|\langle\chi(T)|\psi(T)\rangle|^2=|\langle\chi(t)|\psi(t)\rangle|^2$$

It is useful to define the component $|\xi\rangle$ of the backward-evolved target state $|\chi(T)\rangle$, which is orthogonal to the instantaneous state $|\psi(t)\rangle$:

$$|\xi\rangle=\frac{|\chi\rangle\langle\chi|-F}{\sqrt{F(1-F)}}|\psi\rangle$$

The norm of this component reduces at a rate:

$$Q(t)=\mathrm{Re}\langle\xi(t)|\dot{\psi}(t)\rangle=\mathrm{Im}\langle\xi(t)|\hat{\mathcal{H}}(t)|\psi(t)\rangle$$

which is denoted the direct Hilbert velocity[9]. As described in ref. 9, the trade-off between process duration and achievable fidelity obeys the relation:

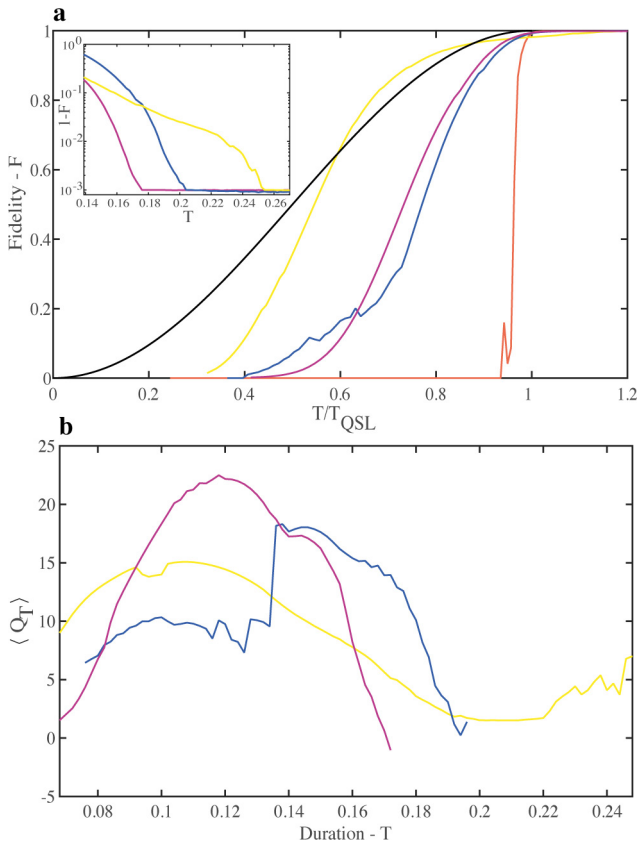
$$\frac{\mathrm{d}F}{\mathrm{d}T}=\frac{2\sqrt{F(1-F)}}{T}\int_0^T Q(t)\mathrm{d}t=2\sqrt{F(1-F)}\,\langle Q\rangle_T \tag{4}$$

where $\langle Q\rangle_T$ is the time average of Q over the process duration. Equation (4) is always true for a uniform extension of time and it is valid for any extension of time for an optimal process[9]. Integrating equation (4) for an optimal clan with vanishing initial fidelity leads to:

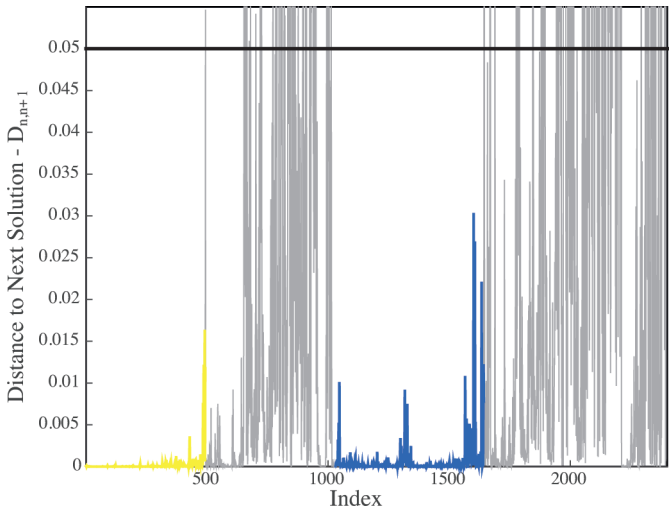
$$F=\sin^2\left(\int_0^T\langle Q\rangle_{T'}\mathrm{d}T'\right) \tag{5}$$

which implies that whenever $\langle Q\rangle_T$ is independent of time, $T_{\mathrm{QSL}}=\pi/2\langle Q\rangle_T$ and for shorter durations, $F(T)=\sin^2\left(\frac{\pi}{2}T/T_{\mathrm{QSL}}\right)$. However, a priori there is no reason that the average direct Hilbert velocity should be constant over an extended range of process durations. $\langle Q\rangle_T$ has been calculated for BringHomeWater for the optimal CHOP paths using the method described above, as shown in Extended Data Fig. 1b. Note that $|\xi\rangle$ and hence $\langle Q\rangle_T$ is only defined for durations where $F<1$. Extended Data Fig. 1b shows that $\langle Q\rangle_T$ is not constant in time, which explains the deviations from $\sin^2\left(\frac{\pi}{2}T/T_{\mathrm{QSL}}\right)$ behaviour in Extended Data Fig. 1a. How the fidelity drops when approaching the QSL for the different strategies can be seen in the inset to Extended Data Fig. 1a.

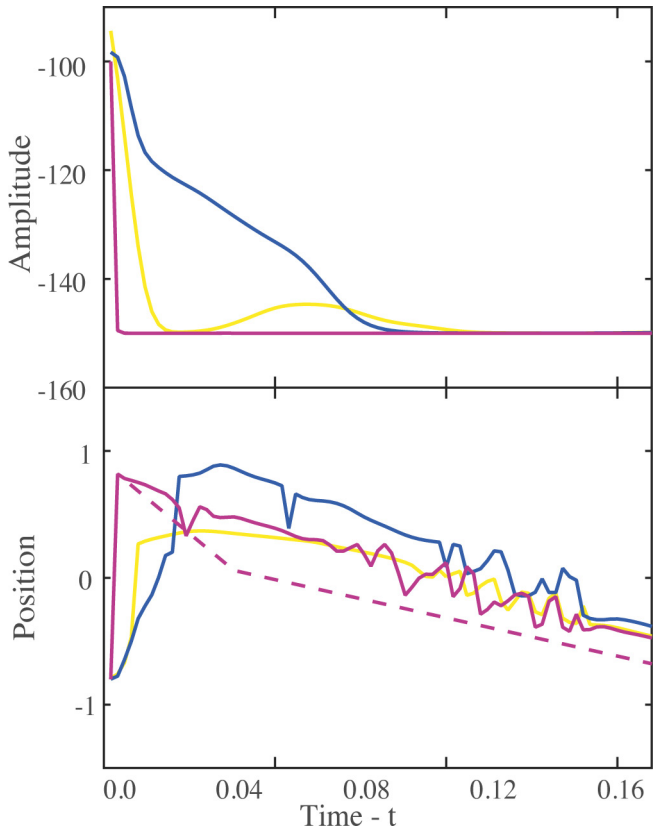
31. Lieberoth, A. *et al.* Getting humans to do quantum optimization—user acquisition, engagement and early results from the citizen cyberscience game Quantum Moves. *Human Comput.* **1,** 221–246 (2014).
32. Sauermann, H. & Franzoni, C. Crowd science user contribution patterns and their implications. *Proc. Natl Acad. Sci. USA* **112,** 679–684 (2015).
33. Calarco, T., Dorner, U., Julienne, P., Williams, C. & Zoller, P. Quantum computations with atoms in optical lattices: marker qubits and molecular interactions. *Phys. Rev. A* **70,** 012306 (2004).
34. Anderlini, M. *et al.* Controlled exchange interaction between pairs of neutral atoms in an optical lattice. *Nature* **448,** 452–456 (2007).
35. Jørgensen, N. B., Bason, M. G. & Sherson, J. F. One- and two-qubit quantum gates using superimposed optical-lattice potentials. *Phys. Rev. A* **89,** 032306 (2014).



Extended Data Figure 1 | Deviations from $\sin^2(\bullet)$ behaviour. Colouring as in Fig. 2. **a**, The fidelity as a function of duration for the optimal families rescaled by the apparent QSL. The T_{QSL} is found for each solution by fitting $\sin^2(aT+b)$ where a and b are fitting parameters. The black line shows $\sin^2(\pi/2T)$ for reference. The inset shows the infidelity $(1-F)$ as a function of duration close to the $T_{\mathrm{QSL}}^{\mathrm{num}}$. **b**, The direct Hilbert velocity $\langle Q\rangle_T$ for the best solutions found by the HILO (purple), tunnelling (yellow) and shovelling (blue) clans respectively. Note that $\langle Q\rangle_T$ is only defined for durations with $F<1$, so the curves end at different durations. The varying direct Hilbert velocity explains the deviation from $\sin^2\left(\frac{\pi}{2}T/\,T_{\mathrm{QSL}}\right)$ in **a**.



Extended Data Figure 2 | The reachability plot for Fig. 3a. The distance between subsequent solutions as calculated by equation (1) in a list sorted according to the pairwise distance between solutions. The index is the position of the solutions in the list. Valleys identify closely spaced solutions, denoted clans. The valleys corresponding to tunnelling and shovelling clans are marked with yellow and blue respectively. The black line marks the threshold for the distance between consecutive solutions in a clan at 0.05.



Extended Data Figure 3 | Solutions from CHOP and HILO. The amplitude and the position of the tweezer as a function of time for the best player solutions in the tunnelling (yellow) and shovelling (blue) clans and the best HILO (purple). The dashed purple line shows the initial seed used by the best HILO solution (note that the dashed and solid purple lines in the top panel overlap). The total duration is $T = 0.15$.

LETTER

doi:10.1038/nature17167

Direct observation of dynamic shear jamming in dense suspensions

Ivo R. Peters[1]†, Sayantan Majumdar[1] & Heinrich M. Jaeger[1]

Liquid-like at rest, dense suspensions of hard particles can undergo striking transformations in behaviour when agitated or sheared[1]. These phenomena include solidification during rapid impact[2,3], as well as strong shear thickening characterized by discontinuous, orders-of-magnitude increases in suspension viscosity[4–8]. Much of this highly non-Newtonian behaviour has recently been interpreted within the framework of a jamming transition. However, although jamming indeed induces solid-like rigidity[9–11], even a strongly shear-thickened state still flows and thus cannot be fully jammed[12,13]. Furthermore, although suspensions are incompressible, the onset of rigidity in the standard jamming scenario requires an increase in particle density[9,10,14]. Finally, whereas shear thickening occurs in the steady state, impact-induced solidification is transient[2,15–17]. As a result, it has remained unclear how these dense suspension phenomena are related and how they are connected to jamming. Here we resolve this by systematically exploring both the steady-state and transient regimes with the same experimental system. We demonstrate that a fully jammed, solid-like state can be reached without compression and instead purely with shear, as recently proposed for dry granular systems[18,19]. This state is created by transient shear-jamming fronts, which we track directly. We also show that shear stress, rather than shear rate, is the key control parameter. From these findings we map out a state diagram with particle density and shear stress as variables. We identify discontinuous shear thickening with a marginally jammed regime just below the onset of full, solid-like jamming[20]. This state diagram provides a unifying framework, compatible with prior experimental and simulation results on dense suspensions, that connects steady-state and transient behaviour in terms of a dynamic shear-jamming process.

Jamming transitions transform fluid-like particle systems into amorphous solids with finite yield stress when the particle packing fraction ϕ increases beyond a critical value, ϕ_J. In the standard scenario[10], the jammed state is reached via isotropic compression, and for frictionless particle interactions the jammed system will weaken and eventually unjam when shear stress is applied. In suspensions, on the other hand, shear can have the opposite role, by inducing viscosity increases and even solidification. Here the idea has been that shear reorganizes particles into anisotropic configurations that form large clusters and potentially a load-bearing network. With frictionless, purely hydrodynamic interactions between suspended particles, such 'hydroclusters'[21,22] can, however, give rise only to mild increases in viscosity (continuous shear thickening)[6,13,23]. Frictional contacts are required to produce the large jumps in viscosity associated with strong, discontinuous shear thickening (DST)[6,7,13,24,25].

How shear can convert isotropic unjammed particle configurations below ϕ_J into anisotropic jammed configurations was shown explicitly for dry granular systems by Bi *et al.*[18] if the particle interactions are frictional and by Kumar and Luding for the frictionless case[19]. This introduced to the original jamming phase diagram a new regime of shear jamming, which subsequently has been adopted as a candidate mechanism for DST[6,8,13]. However, key conceptual as well as experimental questions have remained open. In particular, DST occurs under steady-state shearing conditions and thus cannot involve an actual solid-like, jammed response. Importantly, the onset stress for DST is known to be essentially independent of packing fraction[5], while the predicted onset stress for shear jamming decreases and reaches zero as ϕ_J is approached[18]. Furthermore, direct observations of solid-like shear jamming under controlled conditions have so far been in quasi-static systems of dry grains[18,19]. In other situations where suspensions appear to jam fully—such as in impact-induced solidification—the conditions are typically more complex and in principle can involve shear as well as compression[2,3,17].

To sort this out requires experiments on dense suspensions where solidification fronts can be generated solely by shear and where also DST can be observed. We achieve this here by using a Couette-type geometry. Our results show how solidification fronts produce a solid-like shear-jammed state. This state is qualitatively different from DST in that it exhibits a yield stress, which we demonstrate explicitly by showing how it prevents weights from sinking into the suspension. Our findings suggest that DST does not correspond to a jammed but instead to a fragile state, which exhibits intermittent flow that can be thought of as a precursor to shear jamming and exhibits behaviour of marginal material that is neither freely flowing nor fully jammed[20].

The experiments were performed using a large gap between the inner and outer wall of the cylindrical cell filled with the suspension, which allowed us directly to observe both the transient and steady-state velocity profiles by imaging from above with a high-speed camera (Fig. 1). The suspensions consisted of density-matched solutions of water, glycerol and CsCl, mixed with cornstarch, at packing fractions ϕ from 0.43 by volume up to values close to the isotropic ϕ_J for this material. A key feature of our experiment was that by rotating the inner cylinder in this Couette-type geometry we were applying only shear, thus eliminating any compression that could result in a substantial increase in packing fraction.

Starting with the suspension at rest, sudden rotation of the inner cylinder initiates jamming fronts that propagate radially outward, converting fluid-like into solid-like material, similar to what is observed for impact normal to a free suspension surface[2,17]. To illustrate the effect of rate dependence on this transient response we show in Fig. 1 the evolution of the velocity field for two different rotation speeds. We find strikingly different behaviour between slow and fast driving of the inner cylinder. For a slow driving speed (Fig. 1, top), we observe a velocity profile that behaves in a diffusive manner, reminiscent of that of a viscous liquid, where the spreading slows down as time progresses and monotonically approaches an equilibrium profile (which is similar to those measured in ref. 8). At high driving speeds a sharp front develops (Fig. 1, bottom), which rapidly travels radially outward. Behind the front the velocity of the suspension is fairly uniform, while ahead of the front the suspension is still at rest. After the front reaches the outer wall

[1]James Franck Institute, The University of Chicago, Chicago, Illinois 60637, USA.
†Present address: Engineering and the Environment, University of Southampton, Highfield, Southampton SO17 1BJ, UK.

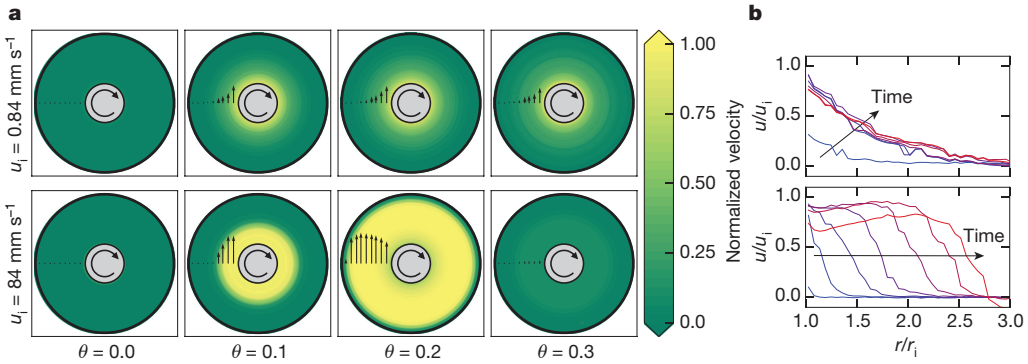


Figure 1 | Transition from viscous response to rapid front propagation. **a**, Evolution of the radial velocity profile (indicated by the arrows and colour), obtained from particle image velocimetry, for two different driving speeds: the bottom row is a hundred times faster than the bottom row. Time is increasing from left to right. Data for both driving speeds are shown for the same amount of shear strain, given by the rotation angle θ. All velocities are normalized by the driving speed u_i at the outer edge of the inner cylinder. Whereas slow driving results in a velocity profile that gradually changes until it reaches a fluid-like steady state, fast driving shows the formation of a travelling jamming front and eventually a solid-like shear-jammed state (bottom row, far right). **b**, Velocity versus radial distance corresponding to the same two driving speeds as in **a**. The radial distance r is normalized by the radius r_i of the inner cylinder. Different curves represent different instances in time, with time increasing from blue to red. The steady-state diffusive profile (top) and the travelling front (bottom) are clearly apparent.

(bottom right in Fig. 1a), the system fully jams while the inner cylinder continues to rotate, creating a plug of solid-like material between the inner and outer cylinder.

When fronts develop, their speed is proportional to the driving speed, as shown in Fig. 2a. Rescaling the x-axis by the driving speed collapses the data (Fig. 2b), indicating that the ratio between driving speed and front propagation speed is independent of the driving speed. The absence of any timescale indicates that the front propagation is resulting only from a critical shear strain, which locally shear-jams the system, not unlike neighbouring gears that engage. Once this critical strain has been reached between any two adjacent radial layers of the suspension, the now shear-jammed portion will be able to strain the still unjammed suspension ahead of it. This process will continue for as long as the system is actively driven and no boundary is reached. The closer the packing density is to the jamming point, the smaller will be the critical strain required to reach a shear-jammed state[18]. This implies that the propagation speed will increase with packing fraction, as shown explicitly in Fig. 2b.

We can take the connection between front propagation in suspensions and shear jamming of dry granular systems one important step further by accounting for a key feature of the shear-jamming phase diagram of ref. 18, namely that both critical stress and strain approach zero as ϕ approaches ϕ_J. Such vanishing critical strain would, according to our reasoning above, result in a diverging front speed, as an infinitesimal perturbation could immediately propagate through the whole system. The functional form of how the critical strain approaches zero, and consequently how the speed diverges as ϕ_J is approached, is not known. Our data, however, can be approximated by the same functional dependence $u_f/u_i = (\phi_0/(\phi_J - \phi_0))^\alpha$ derived for speeds of compression-induced fronts in dry granular materials close to the jamming point[26]. Here u_f is the front speed and u_i the driving, or impact, speed. In the original expression for compression fronts, the exponent $\alpha = 1$. Treating α and ϕ_J as fitting parameters, we find $\alpha = 1.0 \pm 0.1$ also for shear-jamming fronts, and we obtain an estimated $\phi_J \approx 0.56$.

Figure 2d shows that there is a critical driving speed beyond which the torque response jumps by several orders of magnitude, much like in a discontinuous shear thickening (DST) transition. Right at the transition (red curve), the torque wanders off, which shows that the transition itself cannot be resolved in a rate-controlled manner. We therefore turn to stress-controlled measurements to address the central remaining question, namely how this shear-jammed state relates

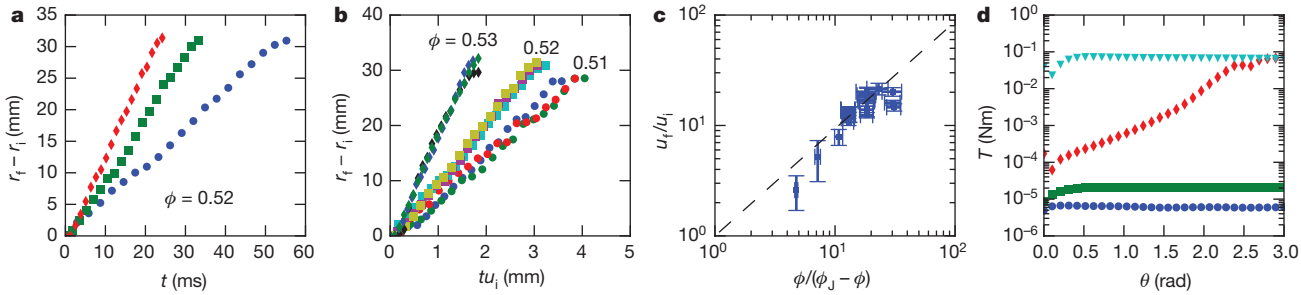


Figure 2 | Influence of driving speed and packing fraction. **a**, Radial position of the shear front (defined as the position where $u = u_i/2$ as a function of time for different driving speeds (blue circles, $u_i = 59$ mm s^{-1}; green squares, $u_i = 92$ mm s^{-1}; red diamonds, $u_i = 126$ mm s^{-1}), for packing fraction $\phi = 0.52$. Here r_i is the radius of the inner cylinder. **b**, Front positions normalized by driving speed (different driving speeds indicated by different colours) for different packing fractions (circles, $\phi = 0.51$; squares, $\phi = 0.52$; diamonds, $\phi = 0.53$). The collapse of data shows that the jamming front propagates due to a critical shear strain. **c**, Front speed $u_f = dr_f/dt$ over driving speed ratio versus packing fraction. The black dashed line has a slope $\alpha = 1.0$, showing that the speed ratio is approximately proportional to $\phi/(\phi_J - \phi)$. Error bars indicate the standard deviation of 7–19 repeated experiments. **d**, Measured torque T for different, fixed driving speeds. The blue circles (84 μm s^{-1}) and green squares (0.84 mm s^{-1}) correspond to the behaviour shown in the upper half of Fig. 1. Cyan triangles (84 mm s^{-1}) correspond to the front formation shown in the bottom row of Fig. 1. The torque is 3–4 orders of magnitude larger, signalling that the front has reached the container wall and a fully jammed, solid-like state has been established (the torque asymptotes because of slip at the wall of the inner, driven cylinder). Red diamonds (8.4 mm s^{-1}) show the situation right at the transition to front formation. Here the suspension goes through the DST regime into a fully shear-jammed state as the applied torque increases while the driving speed is kept constant.

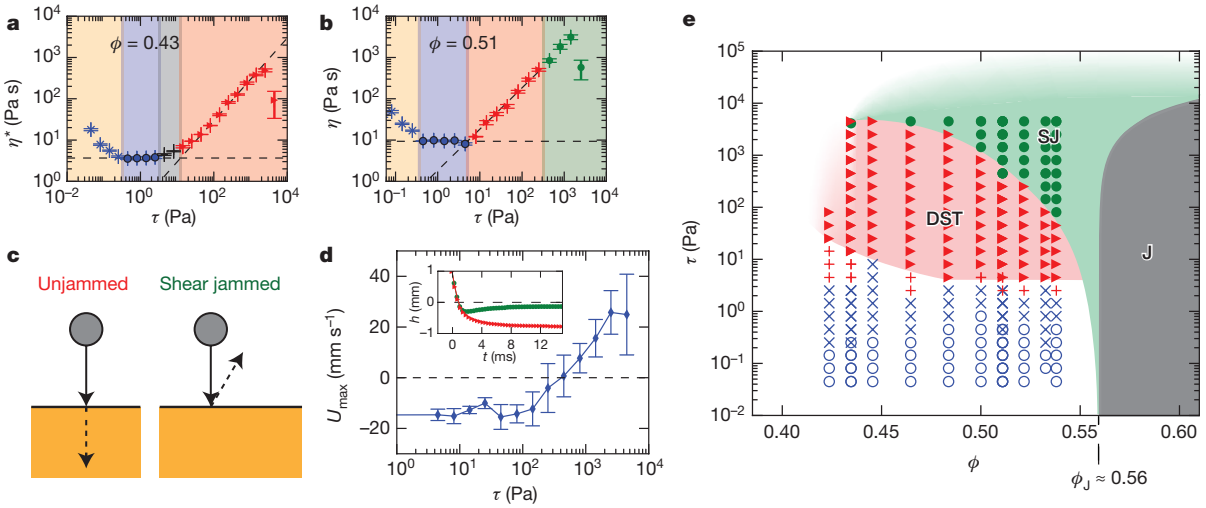


Figure 3 | Transitions between the different states of the suspension. **a**, **b**, Viscosity–stress curves for two different packing fractions, with different states indicated by different background colours: shear thinning (beige), Newtonian (blue), weak shear thickening (grey), discontinuous shear thickening (red), and shear-jammed (green). At high packing fractions, the weaker shear thickening state becomes negligible. Error bars are standard deviations of the temporal fluctuations in viscosity. Note that both data sets show one data point where slip occurs on the inner cylinder (on the far right), which results in a much lower apparent viscosity. Clearly, these data points are not suitable to use to determine an effective viscosity. **c**, Cartoon of the impact experiment. **d**, Transition curve for the crossover from DST to shear jamming, showing the maximum observed upward velocity U_{max} of the sphere after impact as a function of applied shear stress. Error bars give the standard deviation obtained from 5–10 repeated experiments. Inset, two experimental trajectories of spheres impacting on the suspension under two different shear stresses, showing a slowly sinking sphere in the DST regime (red curve) and rebound in the shear-jammed regime (green curve). Here, h is the vertical distance between the bottom of the sphere and a fixed reference point in the high-speed movies (see for example Supplementary Video 1). **e**, State diagram showing the different states of the suspension: shear thinning (blue circles), Newtonian (blue crosses), weaker shear thickening (red crosses), discontinuous shear thickening (red triangles) and shear-jammed (green dots). The shaded areas correspond to the original shear jamming state diagram[18], with isotropic jammed (J) and anisotropic shear-jammed (SJ) regimes. The regime of fragile states is here identified with the DST regime. The regime of shear thinning or Newtonian behaviour at stresses below the onset of DST does not exist in dry granular systems, and in suspensions corresponds to conditions where stresses are too low to allow frictional particle–particle contacts[6,7].

to the well-established steady-state properties of dense suspensions[27]. In previous work, the transient and steady-state behaviours were typically investigated in different systems and by very different methods, since fast transient responses involving front velocities of a few metres per second (Fig. 2a) are difficult to track with standard rheological experiments. In our wide-gap Couette cell, however, we can readily investigate both regimes.

Figure 3a and b shows the steady-state suspension viscosity η^* versus applied shear stress τ for two of the packing fractions investigated. Note that η^* is the apparent viscosity, defined as the ratio of shear stress to shear rate (see Methods). From the slope of the curves in plots like these we identify the following behaviours: Newtonian, shear thinning, weak shear thickening, and DST. Specifically, the DST regime is identified in the most strict sense, that is, by a linear or stronger increase of η^* with τ. We determine the onset of DST by a linear fit to the DST regime on log–log plots such as in Fig. 3a, and find the stress value at which it intersects a fit to the Newtonian regime, where the viscosity is constant, independent of applied stress.

At low ϕ around 0.43, the typical sequence of behaviours with increasing τ within our experimental range moves from shear-thinning via Newtonian to DST (Fig. 3a). At large ϕ around 0.51 (Fig. 3b), the onset of DST is sharper and DST crosses over into a fully shear-jammed state that no longer flows and behaves like a solid. In this shear-jammed regime the inner, rotating cylinder continually slips and the value of η^* is no longer meaningful.

We demonstrate the solid behaviour in the shear-jammed regime explicitly by dropping small steel spheres (diameter 5.0 mm) onto the continuously sheared suspension (Fig. 3c) and tracking their vertical position after they touch the surface (Fig. 3d). As we apply more shear stress to our system, the trajectory of the spheres changes from slowly sinking (unjammed or DST) to rebounding and remaining on the suspension surface for as long as shear stress is applied (shear-jammed). We define the shear-jammed onset stress by the elastic behaviour of the suspension, that is, by the applied shear stress at which the spheres' velocity reverses direction after impact (see Methods).

By performing these experiments across a range of packing fractions below ϕ_J we construct the state diagram shown in Fig. 3e. Its main feature is the delineation between a strongly shear-thickening (DST) regime and a solid-like, shear-jammed regime. As we move towards lower packing fractions, DST is partially replaced by weaker shear thickening. This agrees with prior work that shows DST eventually disappearing at even lower ϕ (ref. 28).

We point out an important difference with previous studies that suggested that DST can be identified directly with the shear-jamming transition[6,8,13]. These studies found that the minimum shear rate for observing DST approaches zero as ϕ approaches ϕ_J. However, the shear-jamming phase diagram, just like the original jamming phase diagram, is controlled by packing density and shear stress, not rate, as shown in Fig. 2c of ref. 18. In fact, other experiments have shown that the onset stress τ_{min} for DST is independent of packing fraction, and, if anything, even increases on the approach to the isotropic jamming point (owing to an emerging yield stress at high packing fractions)[12]. The behaviour of the onset stress for DST would therefore seem to be in direct contradiction to the shear-jamming phase diagram. As seen in Fig. 3e, this issue is easily resolved by recognizing the DST regime as a precursor to the shear-jamming transition, with a lower boundary that is essentially independent of ϕ (see Methods for the slightly different definitions of τ_{min} here and in prior work).

In different experimental geometries, such as parallel plate setups, an upper limit to DST was observed[28], owing to the maximum confinement available from surface tension and given by $\tau_{max} \approx 0.1\sigma/a$. Here σ is the interfacial tension at the free suspension boundary and a is the particle diameter. We find that this limit does not apply with our setup because we see no significant difference when $\sigma \to 0$ (see Methods). It implies that both the flowing DST state and the fully shear-jammed state are highly anisotropic, and so any force network must

lie predominantly within the horizontal plane. Since in the absence of friction such networks would immediately break up as the particles would slide out of plane, we interpret this anisotropy as a signature of the highly frictional particle contacts that enable shear jamming.

This identification of DST is, to the best of our knowledge, consistent with all previous studies and it reframes DST within the context of shear jamming. In particular, the pronounced downward curvature of the upper boundary of DST as ϕ approaches the isotropic jamming point at $\phi_J = 0.56$ follows exactly the qualitative behaviour pointed out by ref. 18 and thus gives the first clear indication of a state diagram that unifies DST with shear jamming.

Our findings suggest the following picture. At low shear stress, particles are not in contact and lubrication forces result in a Newtonian behaviour (or shear thinning in the case of yield stress). Beyond a critical stress, lubrication breaks down, and frictional forces generate a fabric of force chains, resulting in discontinuous shear thickening[6,7,13]. However, rather than being fully jammed, this is a fragile state that intermittently flows and gets stuck. The critical stress required for frictional contacts is independent of packing fraction because it depends only on the microscopic breakdown of lubrication[6,13,24]. With increasing stress, the force network becomes denser until a fully shear-jammed state is reached. This last transition depends strongly on the packing fraction, because the critical shear stress will vanish on the approach to the isotropic jamming point[18].

We point out that these shear-jammed states emerge without any change in packing fraction. This is in contrast with the initial picture developed for solidification under impact[2], which was based on a change in ϕ and needs to be reconsidered within a shear-jamming framework. Finally, our observation of DST states without a transition to shear jamming suggests that, at low packing fractions, fragile states are possible, but will fail at higher stresses before a shear-jammed state is reached. These transitions may possibly be explored in dry granular systems using photoelastic particles.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 11 September 2015; accepted 20 January 2016.
Published online 4 April 2016.

1. Barnes, H. A. Shear-thickening ('dilatancy') in suspensions of nonaggregating solid particles dispersed in Newtonian liquids. *J. Rheol.* **33,** 329–366 (1999).
2. Waitukaitis, S. R. & Jaeger, H. M. Impact-activated solidification of dense suspensions via dynamic jamming fronts. *Nature* **487,** 205–209 (2012).
3. Petel, O. E. *et al.* The effect of particle strength on the ballistic resistance of shear thickening fluids. *Appl. Phys. Lett.* **102,** 064103 (2013).
4. Fall, A., Huang, N., Bertrand, F., Ovarlez, G. & Bonn, D. Shear thickening of cornstarch suspensions as a reentrant jamming transition. *Phys. Rev. Lett.* **100,** 018301 (2008).
5. Brown, E. & Jaeger, H. M. Dynamic jamming point for shear thickening suspensions. *Phys. Rev. Lett.* **103,** 086001 (2009).
6. Seto, R., Mari, R., Morris, J. F. & Denn, M. M. Discontinuous shear thickening of frictional hard-sphere suspensions. *Phys. Rev. Lett.* **111,** 218301 (2013).
7. Wyart, M. & Cates, M. E. Discontinuous shear thickening without inertia in dense non-brownian suspensions. *Phys. Rev. Lett.* **112,** 098302 (2014).
8. Fall, A. *et al.* Macroscopic discontinuous shear thickening versus local shear jamming in cornstarch. *Phys. Rev. Lett.* **114,** 098301 (2015).
9. Cates, M., Wittmer, J., Bouchaud, J.-P. & Claudin, P. Jamming, force chains, and fragile matter. *Phys. Rev. Lett.* **81,** 1841–1844 (1998).
10. Liu, A. J. & Nagel, S. R. Jamming is not just cool any more. *Nature* **396,** 21–22 (1998).
11. Trappe, V., Prasad, V., Cipelletti, L., Segre, P. N. & Weitz, D. A. Jamming phase diagram for attractive particles. *Nature* **411,** 772–775 (2001).
12. Brown, E. *et al.* Generality of shear thickening in dense suspensions. *Nature Mater.* **9,** 220–224 (2010).
13. Mari, R., Seto, R., Morris, J. F. & Denn, M. M. Shear thickening, frictionless and frictional rheologies in non-Brownian suspensions. *J. Rheol.* **58,** 1693–1724 (2014).
14. Keys, A. S., Abate, A. R., Glotzer, S. C. & Durian, D. J. Measurement of growing dynamical length scales and prediction of the jamming transition in a granular material. *Nature Phys.* **3,** 260–264 (2007).
15. Liu, B., Shelley, M. & Zhang, J. Focused force transmission through an aqueous suspension of granules. *Phys. Rev. Lett.* **105,** 188301 (2010).
16. von Kann, S., Snoeijer, J., Lohse, D. & van der Meer, D. Nonmonotonic settling of a sphere in a cornstarch suspension. *Phys. Rev. E* **84,** 060401 (2011).
17. Peters, I. R. & Jaeger, H. M. Quasi-2D dynamic jamming in cornstarch suspensions: visualization and force measurements. *Soft Matter* **10,** 6564–6570 (2014).
18. Bi, D., Zhang, J., Chakraborty, B. & Behringer, R. P. Jamming by shear. *Nature* **480,** 355–358 (2011).
19. Kumar, N. & Luding, S. Memory of jamming—multiscale flow in soft and granular matter. *Granular Matter* http://dx.doi.org/10.1007/s10035-016-0624-2 (in the press); preprint at http://arxiv.org/abs/1407.6167 (2015).
20. Vitelli, V. & van Hecke, M. Marginal matters. *Nature* **480,** 325–326 (2011).
21. Wagner, N. J. & Brady, J. F. Shear thickening in colloidal dispersions. *Phys. Today* **62,** 27–32 (2009).
22. Cheng, X., McCoy, J. H., Israelachvili, J. N. & Cohen, I. Imaging the microscopic structure of shear thinning and thickening colloidal suspensions. *Science* **333,** 1276–1279 (2011).
23. Lin, N. Y. C. *et al.* Hydrodynamic and contact contributions to continuous shear thickening in colloidal suspensions. *Phys. Rev. Lett.* **115,** 228304 (2015).
24. Fernandez, N. *et al.* Microscopic mechanism for shear thickening of non-Brownian suspensions. *Phys. Rev. Lett.* **111,** 108301 (2013).
25. Guy, B. M., Hermes, M. & Poon, W. C. K. Towards a unified description of the rheology of hard-particle suspensions. *Phys. Rev. Lett.* **115,** 088304 (2015).
26. Waitukaitis, S. R., Roth, L. K., Vitelli, V. & Jaeger, H. M. Dynamic jamming fronts. *Europhys. Lett.* **102,** 44001 (2013).
27. Brown, E. & Jaeger, H. M. Shear thickening in concentrated suspensions: phenomenology, mechanisms and relations to jamming. *Rep. Prog. Phys.* **77,** 046602 (2014).
28. Brown, E. & Jaeger, H. M. The role of dilation and confining stresses in shear thickening of dense suspensions. *J. Rheol.* **56,** 875–923 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank E. Brown, E. Han, N. James, S. Mukhopadhyay and Q. Xu for discussions. This work was supported by the US Army Research Office through grant W911NF-12-1-0182 and the Chicago Materials Research Science and Engineering Center, which is funded by the NSF through grant DMR-1420709. S.M. acknowledges support through a Kadanoff-Rice fellowship.

Author Contributions I.R.P., S.M. and H.M.J. conceived the project. I.R.P. and S.M. performed the experiments and the analysis. I.R.P. and H.M.J. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to I.R.P. (i.r.peters@soton.ac.uk).

METHODS

Experimental setup. We used an Anton Paar MCR301 rheometer with a Couette-type cell consisting of two concentric cylinders where the outer cylinder (cup) was a glass container with inner diameter of 98 mm and height of 42 mm. The cup was filled just below the rim with the suspension. The inner cylinder (bob) of diameter of 26.7 mm was controlled by the rheometer and was partially submerged in the suspension. We verified that the measurements were not influenced by the bottom of the cup by performing experiments at different insertion depths of the bob. In addition, we performed tests where we effectively removed the solid bottom boundary by using a layer of Fluorinert (FC-3283, 3 M) liquid on which the suspension floated. At high stresses, we observed slip between the surface of the bob and the suspension. We also performed experiments with a roughened cylinder surface (using sandpaper), but this still resulted in slip at the bob surface, and had no significant influence on our measurements.

Transient behaviour. For these experiments the rheometer was set to rotate at constant speed. The time to accelerate from rest was always much smaller than the time during which we observed the transient response of the system. The experiments were recorded using a high-speed camera (Phantom V9), imaging at rates up to 4,500 frames per second, depending on the driving speed. Tracer particles (ground black pepper) sprinkled on the surface of the cornstarch suspension allowed us to perform particle image velocimetry and obtain time-resolved velocity fields (for example, Fig. 1 and Supplementary Video 2). In addition, we measured the torque applied by the rheometer.

Steady-state behaviour. These experiments were performed by setting a constant torque and measuring the rotation speed as a function of time. We repeated this for a range of torques to obtain viscosity–stress curves. We calculated the stress as $\tau = T/(2\pi R_i^2 h)$, with T the torque, R_i the radius of the bob and h the submerged height of the bob. The local shear rate $\dot{\gamma}$ is a priori unknown (an example of the flow profile at low stress is given in the top row of Fig. 1b). We define an average shear rate via the relation $\dot{\gamma} = \omega(R_o^2 + R_i^2)/(R_o^2 - R_i^2)$ in order to calculate the apparent viscosity $\eta^*/\dot{\gamma}$, with ω the angular rotation rate and R_o the radius of the outer cylinder. Such apparent viscosity corresponds to the true viscosity only for a linear, Newtonian suspension. To calculate the local apparent viscosity everywhere in the system, stress gradients would need to be taken into account in order to obtain the local stress. However, the current method suffices to identify transitions between different states of the suspension with changes in applied shear stress. From the resulting curves of viscosity versus shear stress we define the onset of DST as the intersection of fits to Newtonian and DST regimes as described in the main text. Prior work defined the onset of general (weak or strong) shear thickening via τ_{min}, the minimum in viscosity–stress curves before the start of thickening. Taking this as the criterion in Fig. 3c, we see that even at lower packing fractions the onset stress for shear thickening remains independent of ϕ. To test for the upper stress limit $\tau_{max} \approx 0.1\sigma/a$, identified as the stress scale above which DST is no longer observable in parallel-plate geometries because confinement breaks down, we performed experiments in which we flooded the top free surface of our suspension with a liquid that is miscible with the suspending liquid, a procedure which effectively sets the interfacial tension to zero. In this flooded system, fully shear-jammed, solid-like behaviour was still observed, and stress levels exceeding τ_{max} could easily be achieved. This demonstrates that suspensions driven into a frictional jammed state by (horizontal) shear do not develop notable out-of-plane (vertical) stresses that would need to be balanced by confinement via boundaries or interfacial tension.

Determining the onset of shear jamming. For the impact experiments we released steel spheres with diameter 5 mm and mass 0.51 g from a height of 95 mm; see Supplementary Video 1. An electromagnetic release mechanism was used to ensure reproducible impact conditions. The impacts were recorded with the high-speed camera described above, and analysed using a particle-tracking algorithm. Each experiment was performed up to ten times to obtain a distribution of responses for every combination of packing fraction and shear stress. We note that under our experimental conditions the impacting sphere merely probes the state of the suspension and does not cause it to undergo a transition into a jammed configuration.

To determine whether a system is jammed, we ideally need to test whether the system has a yield stress. Here we use an elastic response (rebound) as a proxy to having a yield stress, because a rebound can be determined more precisely and unambiguously in experiments. We do, however, point out that also for longer times the yield stress behaviour is apparent. An example of this can be observed in Supplementary Video 1 and Extended Data Fig. 2, where a sphere sits on the surface of a suspension for several seconds, until the supplied stress is turned off.

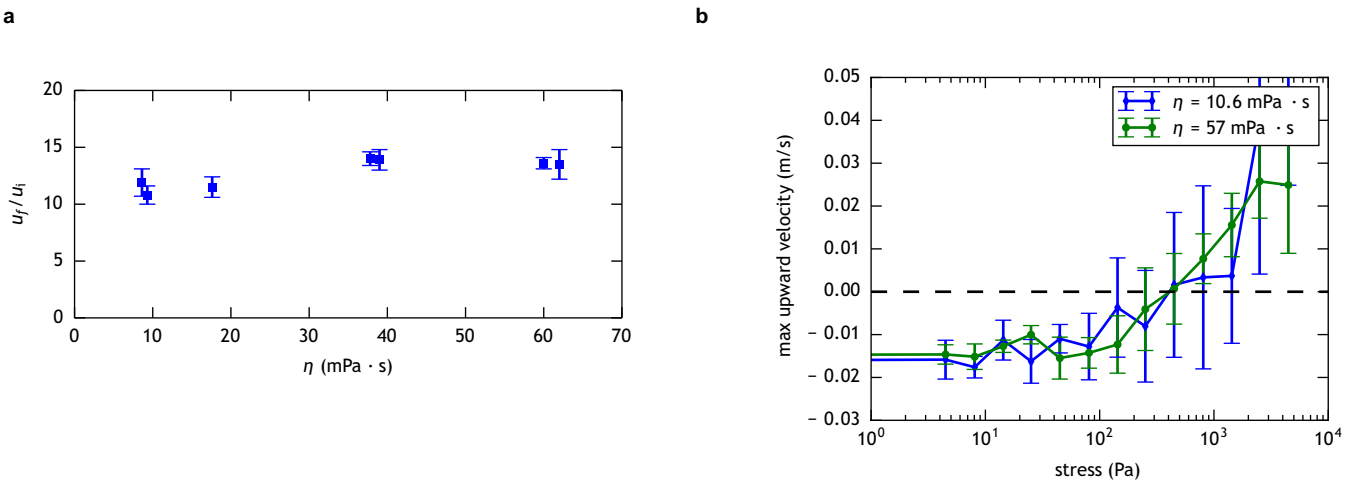
Suspension packing fraction. The suspensions consisted of cornstarch dispersed in a density-matched solution of water, glycerol, and CsCl. The cornstarch was stored at controlled temperature (22.8 ± 0.3 °C) and relative humidity (51 ± 2%). For each solution, we carefully measured the density and liquid viscosity. To calculate the packing fraction of the suspensions, we took into account the porosity of the cornstarch particles. In considering jamming, the relevant quantity is the amount of free interstitial volume available for particle rearrangement. If some of the interstitial liquid penetrates into porous particles the available free volume shrinks. Therefore, a meaningful parameter is the effective packing fraction, which considers the interior to be inaccessible to other particles. Assuming that each particle has a volume fraction $\lambda = 0.3$ (see, for example, refs 29 and 30) of pore space, we can write $\phi = (1 + \lambda)\phi_v$, where ϕ_v is the material packing fraction without accounting for pore space, and ϕ is the packing fraction we quote in this paper. The value of λ is an estimate, but it adjusts only the absolute values of ϕ and has no qualitative influence on our results. To calculate ϕ_v, we take into account the moisture content β, which the cornstarch has absorbed from the environment. Assuming that this moisture content is pure water, we calculate the packing fraction as follows:

$$\phi_v = \frac{(1-\beta)m_{cs}/\rho_{cs}}{(1-\beta)m_{cs}/\rho_{cs} + m_l/\rho_l + \beta m_{cs}/\rho_w} \quad (1)$$

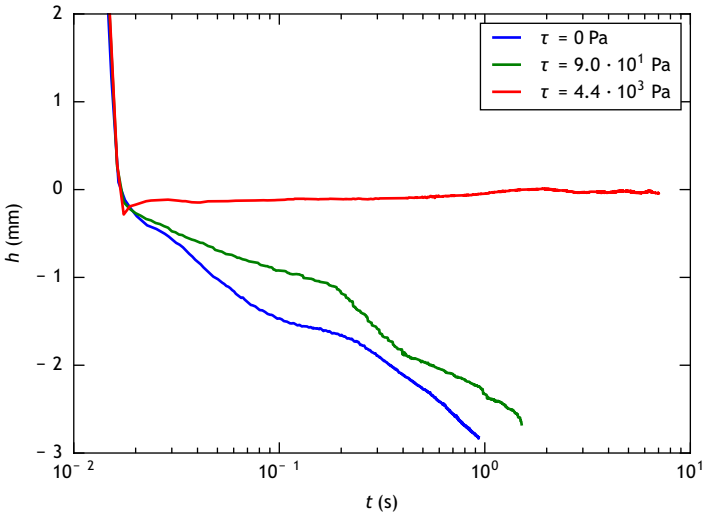
with $\rho_{cs} = 1.62 \times 10^3$ kg m^{-3} the density of the dry cornstarch, ρ_l and m_l the density and mass of the water/glycerol/CsCl solution, m_{cs} is the mass of the cornstarch including the moisture content, and ρ_w is the density of pure water. The moisture content is estimated to be $\beta = 0.1$, but small variations will result in small variations in the packing fraction, which influence the experiments noticeably only when very close to the isotropic jamming point ϕ_I. We found that for experiments performed within a relatively short time frame (a few days), we can assume a constant value for β. The value for β can be determined by performing a set of experiments at different packing fractions and determining the jamming packing fraction, which should be the same for all experiments. Sets of experiments that were performed when the moisture content had changed could be matched to the existing data sets by adjusting β accordingly.

Influence of solvent viscosity. The ratio between propagation speed of the jamming fronts u_f and the driving speed u_i depend on the packing fraction of the suspension. Because the jamming front is a result of frictional interaction between the particles, this ratio u_f/u_i is expected to be independent of the solvent viscosity if the shear stress is high enough[2,31]. To test this we performed a number of experiments in which we kept the packing fraction constant, and changed only the solvent viscosity. The results are shown in Extended Data Fig. 1a, where we plot the ratio u_f/u_i as a function of the solvent viscosity. We similarly tested the influence of solvent viscosity on the onset of bouncing motion for the impact experiments. For this, we used two different viscosities, as can be seen in Extended Data Fig. 1b. Note from the error bars that in this case the data are less noisy for the more viscous fluid (green data points), which gives a more accurate determination of the onset for bouncing.

29. Sair, L. & Fetzer, W. R. Water sorption by starches. *Ind. Eng. Chem.* **36,** 205–208 (1944).
30. Hellman, N. N. & Melvin, E. H. Surface area of starch and its role in water sorption. *J. Am. Chem. Soc.* **72,** 5186–5188 (1950).
31. Waitukaitis, S. R. *Impact-Activated Solidification of Cornstarch and Water Suspensions* 51–56 (Springer, 2014).



Extended Data Figure 1 | Influence of viscosity on front speed and shear jamming transition. a, Ratio between front speed and driving speed as a function of the solvent viscosity. Error bars are standard deviations of 7–19 repeated experiments at different driving speeds u_i. **b**, Bouncing transition curves for two different solvent viscosities. The horizontal axis gives the applied shear stress, the vertical axis the maximum upward velocity we observed from the trajectory of the impacting sphere. Error bars are standard deviations of 5–10 repeated experiments.



Extended Data Figure 2 | Long-time behaviour of shear-jammed and unjammed suspension. Experimental trajectories showing the long-time behaviour of spheres impacting the suspension under three different shear stresses. Zero shear stress (blue curve) and a stress of 90 Pa (DST, green curve) both show a slowly sinking sphere. The shear-jammed state (4,400 Pa, red curve) shows a rebound followed by yield stress behaviour, shown by the sphere not sinking in.

LETTER

doi:10.1038/nature17438

Asymmetric catalytic formation of quaternary carbons by iminium ion trapping of radicals

John J. Murphy[1]*, David Bastida[1]*, Suva Paria[1], Maurizio Fagnoni[2] & Paolo Melchiorre[1,3]

An important goal of modern organic chemistry is to develop new catalytic strategies for enantioselective carbon–carbon bond formation that can be used to generate quaternary stereogenic centres. Whereas considerable advances have been achieved by exploiting polar reactivity[1], radical transformations have been far less successful[2]. This is despite the fact that open-shell intermediates are intrinsically primed for connecting structurally congested carbons, as their reactivity is only marginally affected by steric factors[3]. Here we show how the combination of photoredox[4] and asymmetric organic catalysis[5] enables enantioselective radical conjugate additions to β,β-disubstituted cyclic enones to obtain quaternary carbon stereocentres with high fidelity. Critical to our success was the design of a chiral organic catalyst, containing a redox-active carbazole moiety, that drives the formation of iminium ions and the stereoselective trapping of photochemically generated carbon-centred radicals by means of an electron-relay mechanism. We demonstrate the generality of this organocatalytic radical-trapping strategy with two sets of open-shell intermediates, formed through unrelated light-triggered pathways from readily available substrates and photoredox catalysts—this method represents the application of iminium ion activation[6] (a successful catalytic strategy for enantioselective polar chemistry) within the realm of radical reactivity.

Organic chemists generally rely on polar reactivity to address the challenge of forming quaternary carbon stereocentres in a catalytic enantioselective fashion[1]. Of the stereoselective methods available, metal-catalysed conjugate addition of organometallic nucleophilic species to trisubstituted unsaturated carbonyl substrates has recently emerged as a powerful technology[7–11] (Fig. 1a). These additions are reliable processes, but they generally require controlled reaction conditions and preformed organometallic reagents[7–10]. In contrast, there has been limited success in developing analogous transformations with nucleophilic carbon-centred radicals. Although a few examples of metal-catalysed enantioselective radical conjugate additions (RCAs) have been reported[12–15], none of these approaches provide for the formation of sterically demanding quaternary carbons. The work we report here was prompted by the desire to address this gap in catalytic enantioselective methodology.

Our initial motivation stems from the notion that, because of the long incipient carbon–carbon bond in the early transition state[16], additions of radicals to electron-deficient olefins are rather insensitive to steric hindrance[3]. This makes radical reactivity particularly suited to connecting structurally complex carbon fragments while forging quaternary carbons[17]. We also recognized that the emerging field of photoredox catalysis[4] had recently provided an effective way of generating radicals from bench-stable precursors and under mild conditions. As a result, novel transformations have been invented that capitalize upon non-traditional open-shell mechanisms[18]. We sought to combine this effective radical generation strategy, which does not require pre-functionalized reagents, with a suitable chiral catalyst that could drive the stereoselective trapping of photogenerated radicals while forging quaternary stereocentres. If successful, this combination would provide direct access to chiral molecules that could not be synthesized using polar conjugate additions.

We used the iminium ion activation strategy[6] to attack the problem of identifying a suitable chiral catalyst. This chemistry exploits the electrophilic nature of the iminium ion **A** (Fig. 1b), generated upon condensation of chiral amine catalysts and α,β-unsaturated ketones, to facilitate enantioselective conjugate additions of nucleophiles[19]. This catalytic platform has found many applications in the polar domain[5,6]. However, to date, iminium ions **A** have not been used to trap nucleophilic radicals. This is most surprising, given the high tendency of open-shell species to react with electron-deficient olefins[3]. We reasoned that this dearth of applications could stem from the nature of the radical intermediate **B**, generated upon carbon–carbon bond formation (Fig. 1c). Generally, olefinic radical traps are electrically neutral and afford long-lived, neutral radical intermediates. In contrast, radical addition to the cationic iminium ion **A** generates a short-lived, highly reactive α-iminyl radical cation **B**, which, in line with the classical behaviour of radical ions[20], has a high tendency to undergo radical elimination (β-scission)[21] to re-form the more stable iminium ion **A**.

The instability of **B** is the main obstacle to productive RCA to iminium ions (Fig. 1d). We considered the possibility of reducing the radical cation **B** *in situ* to generate the corresponding enamine **C**, which can be hydrolysed in a facile manner to release both the catalyst and the conjugate addition product. From the outset, we identified three design elements as key to realizing this goal. First, the high reactivity of **B** requires a very rapid single electron transfer (SET) reduction. We hypothesized that using a chiral amine catalyst with a redox active, electron-rich moiety ('e^- pool' unit in Fig. 1d) attached would secure a fast, proximity-driven intramolecular reduction of **B**. This idea finds support in the mechanism of electron transfer within biological systems, where even endergonic redox processes can be achieved via electron tunnelling if the redox centres are in close proximity[22]. Second, we needed to identify a rapid process to interrupt a possible equilibrium between **B** and the nascent enamine **C** established by an intramolecular back electron transfer (BET). Since secondary enamines are known to exist mainly as tautomeric electron poor imines **D**[23], the use of a chiral primary amine catalyst potentially offered an efficient mechanism to preclude the BET by triggering a tautomeric equilibrium which converts **C** into **D**. Last, the oxidized centre ('e^- hole' unit in Fig. 1d), arising from the intramolecular SET, had to be long-lived enough to undergo SET reduction from the photoredox catalysts, restoring the redox-active moiety while facilitating productive catalysis. Achieving a high level of stereocontrol further complicated matters.

To test the feasibility of this electron-relay strategy[24], we explored the reaction between β-methyl cyclohexenone **1a** and benzodioxole **2a** (Table 1). We used the photocatalyst tetrabutylammonium

[1]Institute of Chemical Research of Catalonia (ICIQ), Barcelona Institute of Science and Technology, Avda. Països Catalans 16, 43007 Tarragona, Spain. [2]Photogreen Laboratory, Department of Chemistry, University of Pavia, viale Taramelli 12, 27100 Pavia, Italy. [3]Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain.
*These authors contributed equally to this work.

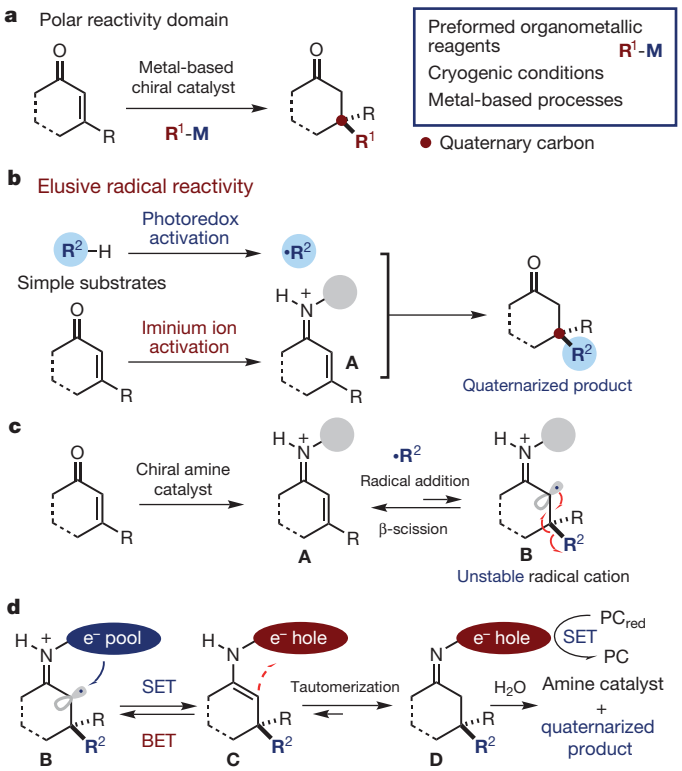
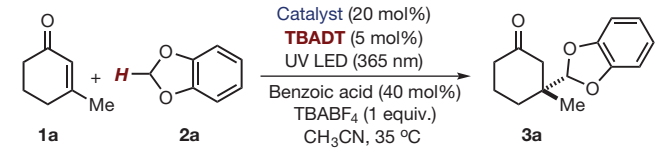
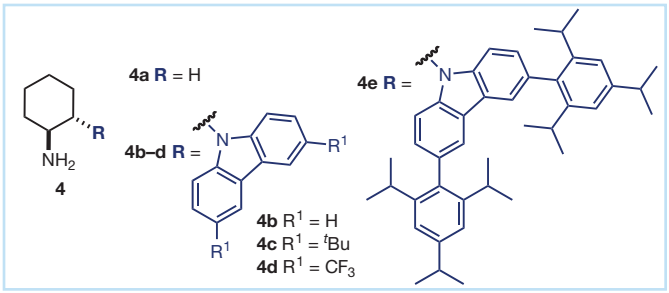


Figure 1 | Conjugate addition technology for forging quaternary stereocentres. a, Established metal-catalysed enantioselective conjugate additions of organometallic reagents (R^1-M) via classical polar pathways. **b**, Design plan for dual photoredox and iminium ion catalysis of radical conjugate additions (RCAs); the grey circle represents the chiral organic catalyst scaffold. **c**, Challenges associated with implementing iminium ion-catalysed conjugate addition of radicals ($\cdot R^2$). **d**, Our electron-relay strategy to rapidly remove the short-lived α-iminyl radical cation (**B**) by intramolecular reduction, and the role of tautomerization to prevent back electron transfer (BET). SET, single electron transfer. PC, photocatalyst; PC_{red}, reduced form of the photocatalyst. The blue ellipse represents an electron rich, reducing moiety, while the red ellipse represents a stable oxidizing species.

Table 1 | Exploratory studies of the feasibility of the electron-relay strategy





Entry	Catalyst	Time (h)	E_p^{ox}/E_p^{red} (**4**) versus Ag/AgCl (V)	**3a** yield (%)	e.e. (%)
1	None	48	NA	Trace	ND
2†	**4a**	48	NA	4	0
3	**4b**	48	+1.15/−1.32	33	82
4†	**4a***	48	NA	5	0
5	**4c**	48	+1.05/−1.28	35	84
6	**4d**	48	+1.51/−1.86	52	77
7	**4e**	48	+1.10/−1.38	46	93
8	**4e**	84	+1.10/−1.38	75	93

Redox potential values (E_p^{ox} and E_p^{red}) describe the electrochemical properties of the redox active carbazole moiety in **4**. NA, not applicable; ND, not determined.
*In combination with 20 mol% of exogenous *N*-cyclohexyl-3,6-di-*tert*-butyl-carbazole.
†Using 40 mol% of trifluoroacetic acid (TFA) instead of benzoic acid.

decatungstate[25] (TBADT, 5 mol%) because, upon light excitation, it can easily generate a nucleophilic carbon-centred radical by homolytically cleaving the methylene C–H bond in **2a**[26] via a hydrogen atom transfer (HAT) mechanism. The experiments were conducted at 35 °C in acetonitrile (CH_3CN) and under irradiation by a single ultraviolet (UV)-light emitting diode (UV LED, $\lambda_{max} = 365$ nm). We observed a negligible racemic background process in the absence of any amine catalyst, which is necessary for realizing a stereoselective process (entry 1). We then focused on identifying a redox-active moiety that, when installed within the chiral primary amine catalyst, could instigate a fast intramolecular reduction of the transient radical cation **B** and thus trigger the entire RCA. We identified carbazole as a suitable scaffold because of (i) its excellent electron-donating capabilities, which would provide the e^- pool unit, and (ii) the high stability of the long-lived carbazole radical cation[27], which makes it a possible e^- hole moiety. These properties form the basis of the wide application of carbazoles in hole-transport materials for light-emitting diodes and photovoltaic cells[28]. The chiral cyclohexylamine scaffold **4b** adorned with the carbazole provided the product **3a** with appreciable yield and stereoselectivity (33% yield, 82% enantiomeric excess (e.e.), entry 3). In consonance with the proposed electron-relay mechanism, the reaction could not be catalysed by cyclohexylamine **4a**, which mimics the catalyst **4b**'s scaffold while lacking the redox-active moiety (entry 2). An equimolar combination of **4a** and exogenous *N*-cyclohexyl-3,6-di-*tert*-butyl-carbazole (20 mol%) also proved unsuitable for catalysis, suggesting the importance of a proximity-driven intramolecular SET process. We then modified the redox properties of the carbazole scaffold by introducing substituents at the 3- and 6-positions. It is known that this substitution pattern can further stabilize the carbazole radical cation[27]. Indeed, we could isolate a bench-stable carbazoliumyl radical cation salt from *N*-cyclohexyl-3,6-di-*tert*-butyl-carbazole upon treatment with $SbCl_5$ (see Supplementary Information). Concurrently, the increased steric hindrance carried the additional benefit of conferring a higher stereocontrol. These considerations explain the high yield and enantioselectivity achieved when using the encumbered primary amine catalyst **4e** (75% yield and 93% e.e., entry 8). Finally, no product formation was detected in the absence of TBADT, catalyst **4e**, or UV light, demonstrating the need for all these components.

We then undertook studies to better investigate the role of the active intermediates in the electron-relay mechanism (Fig. 2a). We could synthesize stable tetrafluoroborate salts of the chiral iminium ion **A-1**, generated upon condensation of catalyst **4c** and substrate **1a**, which were characterized by X-ray single-crystal analysis (Fig. 2b). The unusual stability of the iminium ion **A-1** and the well-defined (*Z*)-configuration of the C=N double bond originate from a stabilizing intramolecular charge transfer π–π interaction between the electron-rich carbazole and the electron-deficient iminium ion. As a result, the measured interatomic separation in the solid state between the carbazole nitrogen and the sp^2 α-carbon of the iminium ion (3.10 Å) is significantly less than the van der Waals distance. This highly organized topology of **A-1**, which NMR spectroscopic analyses confirmed to be also dominant in solution, plays a critical dual role. On the one hand, it governs the stereocontrol of the RCA, since the bulky carbazole unit is positioned in such a way as to effectively shield the diastereotopic *Si* face of the iminium ion, leaving the *Re* face exposed for enantioselective bond formation (*Re* and *Si* are stereochemical descriptors for heterotopic faces). Importantly, the sense of asymmetric induction observed in the model reaction is consistent

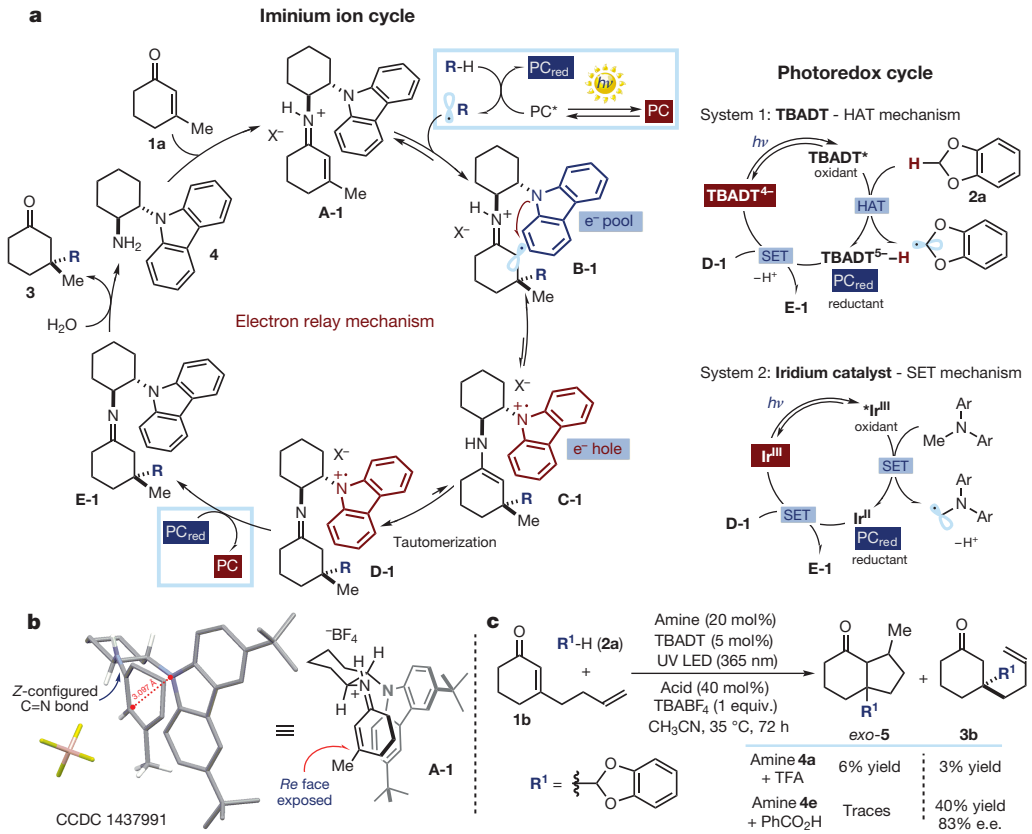


Figure 2 | Proposed mechanism and mechanistic investigations. **a**, Synergistic activities of the iminium ion and the photoredox catalytic cycles to realize the enantioselective RCA to enone **1a**. Upon radical addition to the iminium ion **A-1**, the electron-relay mechanism rapidly reduces the unstable radical cation **B-1** producing a carbazoliumyl radical cation **C-1**, which is prevented from undergoing BET by tautomerization of the secondary enamine to the corresponding imine **D-1**. Regeneration of the photocatalyst (PC) is achieved by reduction of the carbazoliumyl radical cation in **D-1**, while the aminocatalyst **4** is liberated upon hydrolysis of imine **E-1**. **b**, X-ray crystal structure of the carbazole-based iminium ion **A-1**: the distance between the carbazole nitrogen and the sp^2 α-carbon of the cyclohexene moiety is highlighted. **c**, Cyclization experiments indicating that the α-iminyl radical intermediate **B-1** has been bypassed when using the carbazole-based catalyst **4e**.

with this stereochemical model (Fig. 2b). On the other hand, the three-dimensional assembly of **A-1** suggests that the α-iminyl radical cation **B-1**, arising from the radical trapping, is generated in close proximity to the electron-rich carbazole, allowing for a proximity-driven[22] intramolecular reduction. Once the carbazoliumyl radical cation (e^- hole) is generated, the fast tautomerization of the secondary

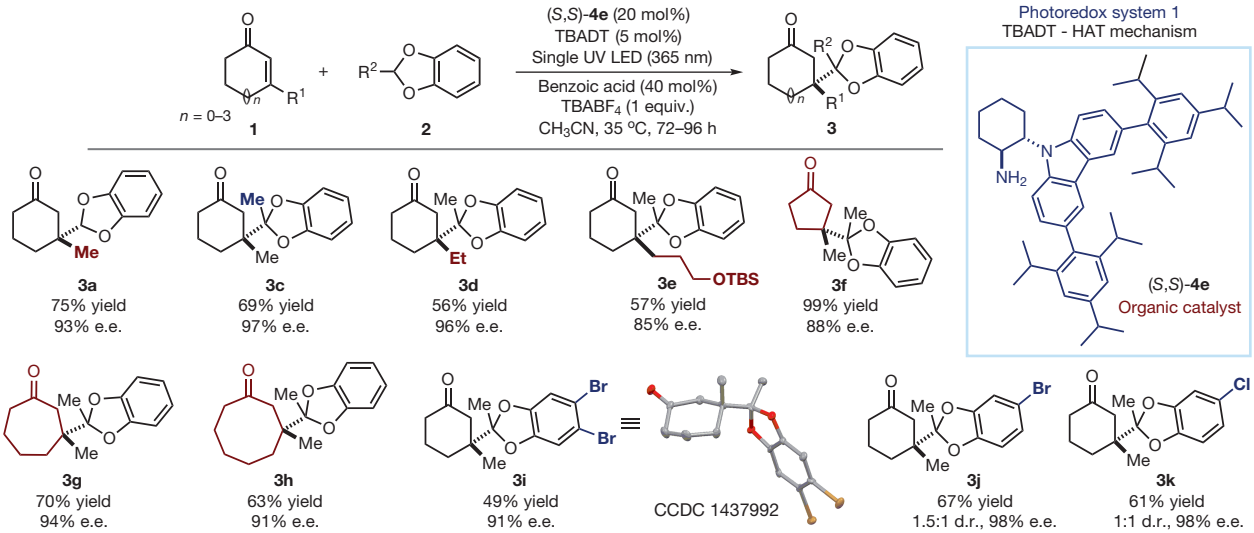


Figure 3 | Substrate scope for the enantioselective trapping of benzodioxole-derived radicals via the dual photoredox organocatalytic strategy. Survey of the cyclic enones **1** and substituted benzodioxoles **2** that can participate in the organocatalytic asymmetric radical conjugate addition (RCA) to forge quaternary stereocentres (as in **3**). Yields and enantiomeric excesses of the isolated products are indicated below each entry (**3a** to **3k**). Details of the TBADT-mediated photoredox cycle to produce carbon-centred radicals from **2** via a HAT mechanism are reported in Fig. 2a. The carbazole-based organic catalyst **4e** is drawn in the boxed inset. TBS, *tert*-butyldimethylsilyl; $TBABF_4$, tetrabutylammonium tetrafluoroborate; d.r., diastereomeric ratio.

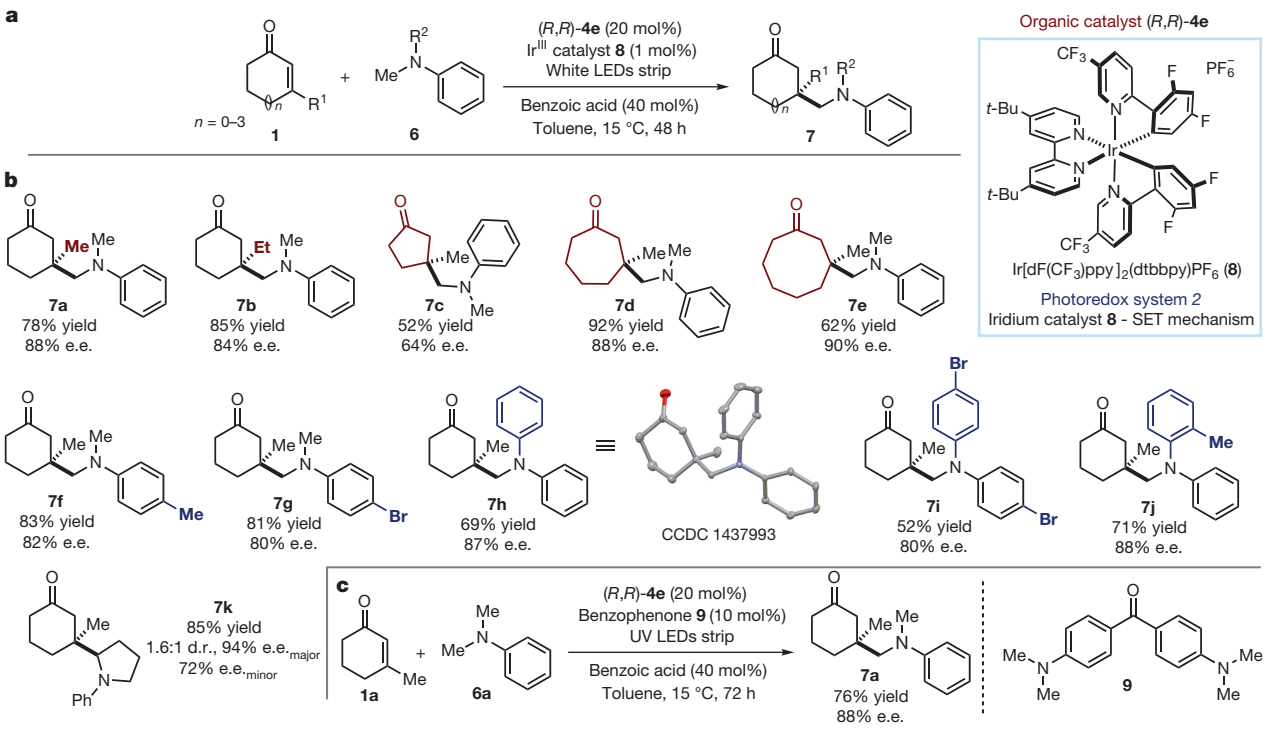


Figure 4 | Substrate scope for the enantioselective trapping of α-amino radicals via the dual photoredox organocatalytic strategy. **a**, The photochemical organocatalytic radical conjugate addition (RCA) developed to forge quaternary stereocentres. **b**, Survey of the cyclic enones **1** and tertiary amines **6** that can participate in the reaction. Yields and enantiomeric excesses (e.e.) of the isolated products **7a–k** are indicated below each entry. Details of the iridium-mediated photoredox cycle to produce carbon-centred radicals via a SET mechanism are reported in Fig. 2a. The structure of the iridium photoredox catalyst **8** is given in the boxed inset. **c**, Fully organocatalytic enantioselective RCA using the benzophenone photocatalyst **9**.

enamine **C-1** to afford the imines **D-1** precludes the BET. At this point, the long-lived radical cation in **D-1** ($E_p^{red} > -1.28$ V versus Ag/Ag$^+$ in CH_3CN, see Table 1) can be reduced by the photocatalyst (half-wave potential $E_{1/2}$ [TBADT^{4-}/TBADT^{5-}−H] = −0.96 V versus Ag/Ag$^+$ in CH_3CN)[29], closing the photoredox cycle (Fig. 2a, right panel). The iminium ion cycle terminates with the imine **E-1** hydrolysis to regenerate the catalyst **4** while liberating the product **3**.

To gain further evidence supporting the electron-relay mechanism, we used the enone **1b**, bearing a β-homoallyl substituent, to trap the radical photogenerated from **2a** (Fig. 2c). The reaction catalysed by cyclohexylamine **4a** provides preferentially the cyclized *exo*-adduct **5** (**5**:**3b** in a 2:1 ratio), demonstrating the propensity of the α-iminyl radicals, emerging from the radical addition, to undergo cyclization with unactivated olefins. In sharp contrast, the process catalysed by the amine **4e** almost exclusively affords the conjugate addition product **3b** (40% yield, 83% e.e., **3b**:**5** in a >10:1 ratio). This result is consonant with a fast redox process, governed by the carbazole-based catalyst, which rapidly reduces the α-iminyl radical cation **B-1** preventing cyclization.

Adopting the optimized conditions described in Table 1, entry 8, we then demonstrated the generality of the RCA by evaluating a variety of cyclic enones **1** and benzodioxoles **2** (Fig. 3). The presence of a methyl substituent at the methylene position of **2** provides the corresponding product **3c**, bearing two adjacent tetrasubstituted carbons, with nearly perfect enantioselectivity. Experiments to probe the scope of the enone component revealed that a wide range of carbocycles and β-olefin substituents are well tolerated. For example, high levels of stereocontrol are achieved with different β-alkyl groups (products **3d**, **e**) and with a diverse range of ring sizes, including cyclopentenyl, cycloheptenyl, and cyclooctenyl architecture (adducts **3f–h**). One limitation is that the presence of an aromatic β-substituent completely inhibits the reaction. As for the benzodioxole substrates **2**, different substituents can be installed at the aromatic ring without compromising the efficiency of the reaction (adducts **3i–k**). Crystals of compound **3i** were suitable for X-ray analysis, which secured the absolute configuration of the products.

We then wondered if the synthetic utility of the amine carbazole catalyst **4e** could be expanded to trap other carbon-centred radicals, formed through an unrelated light-triggered mechanism, while forging quaternary stereocentres. Specifically, we used the commercially available photocatalyst Ir[dF(CF_3)ppy]$_2$(dtbbpy)PF_6 (**8**) which, upon absorption of visible light, can generate α-amino radicals directly from tertiary amines **6** via single electron oxidation[18] (SET mechanism in Fig. 2a). The conjugate addition adducts **7** were provided with high stereoselectivity by using the combination of catalysts (*R*,*R*)-**4e** and **8** while conducting the reactions with enones **1** at 15 °C in toluene and under irradiation by a white LED ($\lambda_{emiss} > 400$ nm) (Fig. 4a). We next explored the scope of both substrates in this dual photoredox organocatalytic strategy. As highlighted in Fig. 4b, cyclic enones of different ring sizes (**7c–e**) and bearing alkyl β-substituents (products **7a**, **b**) are suitable substrates, while both mixed *N*-alkyl-*N*-aryl (adducts **7a**, **f**, **g**) and *N*,*N*-diaryl tertiary amines (**7h–j**) efficiently participated in the RCA. Substituents of different electronic nature were easily accommodated at the aryl *para* (**7f**, **g**, **i**) or *ortho* position (**7j**), while a cyclic amine afforded compound **7k** with high enantiomeric purity, albeit with a 3:2 diastereomeric ratio. For this enantioselective trap of α-amino radicals, we determined a quantum yield of 0.4 ($\lambda = 400$ nm), while Stern–Volmer fluorescence quenching experiments demonstrated that the excited state of the photocatalyst **8** is quenched by the amine **6**. Both experiments are consistent with the electron-relay mechanism depicted in Fig. 2a. Notably, the RCA can be performed without any metal when replacing the photocatalyst **8** with the benzophenone **9**, which can generate the radical acting as an organic photosensitizer[30] (Fig. 4c).

We have developed the first (to our knowledge) catalytic strategy that allows quaternary carbon stereocentres to be obtained with high fidelity using an enantioselective RCA manifold. The approach

requires mild conditions and unfunctionalized substrates, effectively complementing established polar conjugate addition technologies based on preformed organometallic reagents.

Received 23 November 2015; accepted 17 February 2016.

1. Quasdorf, K. W. & Overman, L. E. Catalytic enantioselective synthesis of quaternary carbon stereocentres. *Nature* **516,** 181–191 (2014).
2. Murakata, M., Jono, T., Mizuno, Y. & Hoshino, O. Construction of chiral quaternary carbon centers by catalytic enantioselective radical-mediated allylation of α-iodolactones using allyltributyltin in the presence of a chiral Lewis acid. *J. Am. Chem. Soc.* **119,** 11713–11714 (1997).
3. Fischer, H. & Radom, L. Factors controlling the addition of carbon-centered radicals to alkenes — an experimental and theoretical perspective. *Angew. Chem. Int. Ed.* **40,** 1340–1371 (2001).
4. Schultz, D. M. & Yoon, T. P. Solar synthesis: prospects in visible light photocatalysis. *Science* **343,** 1239176 (2014).
5. MacMillan, D. W. C. The advent and development of organocatalysis. *Nature* **455,** 304–308 (2008).
6. Lelais, G. & MacMillan, D. W. C. Modern strategies in organic catalysis: the advent and development of iminium activation. *Aldrichim. Acta* **39,** 79–87 (2006).
7. Hawner, C. & Alexakis, A. Metal-catalyzed asymmetric conjugate addition reaction: formation of quaternary stereocenters. *Chem. Commun.* **46,** 7295–7306 (2010).
8. Hird, A. W. & Hoveyda, A. H. Catalytic enantioselective alkylations of tetrasubstituted olefins. Synthesis of all-carbon quaternary stereogenic centers through Cu-catalyzed asymmetric conjugate additions of alkylzinc reagents to enones. *J. Am. Chem. Soc.* **127,** 14988–14989 (2005).
9. Shintani, R., Tsutsumi, Y., Nagaosa, M., Nishimura, T. & Hayashi, T. Sodium tetraarylborates as effective nucleophiles in rhodium/diene-catalyzed 1,4-addition to β,β-disubstituted α,β-unsaturated ketones: catalytic asymmetric construction of quaternary carbon stereocenters. *J. Am. Chem. Soc.* **131,** 13588–13589 (2009).
10. Liu, Y., Han, S.-J., Liu, W.-B. & Stoltz, B. M. Catalytic enantioselective construction of quaternary stereocenters: assembly of key building blocks for the synthesis of biologically active molecules. *Acc. Chem. Res.* **48,** 740–751 (2015).
11. Sidera, M., Roth, P. M. C., Maksymowicz, R. M. & Fletcher, S. P. Formation of quaternary centers by copper-catalyzed asymmetric conjugate addition of alkylzirconium reagents. *Angew. Chem. Int. Ed.* **52,** 7995–7999 (2013).
12. Srikanth, G. S. C. & Castle, S. L. Advances in radical conjugate additions. *Tetrahedron* **61,** 10377–10441 (2005).
13. Sibi, M. P., Ji, J., Wu, J. H., Gürtler, S. & Porter, N. A. Chiral Lewis acid catalysis in radical reactions: enantioselective conjugate radical additions. *J. Am. Chem. Soc.* **118,** 9200–9201 (1996).
14. Gansäuer, A., Lauterbach, T., Bluhm, H. & Noltemeyer, M. A catalytic enantioselective electron transfer reaction: titanocene-catalyzed enantioselective formation of radicals from *meso*-epoxides. *Angew. Chem. Int. Ed.* **38,** 2909–2910 (1999).
15. Ruiz Espelt, L., McPherson, I. S., Wiesnsch, E. M. & Yoon, T. P. Enantioselective conjugate additions of α-amino radicals via cooperative photoredox and Lewis acid catalysis. *J. Am. Chem. Soc.* **137,** 2452–2455 (2015).
16. Damm, W. *et al.* Diastereofacial selectivity in reactions of substituted cyclohexyl radicals. An experimental and theoretical study. *J. Am. Chem. Soc.* **114,** 4067–4079 (1992).
17. Schnermann, M. J. & Overman, L. E. A concise synthesis of (−)-Aplyviolene facilitated by a strategic tertiary radical conjugate addition. *Angew. Chem. Int. Ed.* **51,** 9576–9580 (2012).
18. Prier, C. K., Rankic, D. A. & MacMillan, D. W. C. Visible light photoredox catalysis with transition metal complexes: applications in organic synthesis. *Chem. Rev.* **113,** 5322–5363 (2013).
19. Gu, X. *et al.* A general, scalable, organocatalytic nitro-Michael addition to enones: enantioselective access to all-carbon quaternary stereocenters. *Org. Lett.* **17,** 1505–1508 (2015).
20. Poniatowski, A. J. & Floreancig, P. E. in *Carbon-Centered Free Radicals and Radical Cations: Structure, Reactivity, and Dynamics* Ch. 3 (ed. Forbes, M. D. E.) 43–49 (Wiley & Sons, 2010).
21. Jakobsen, H. J., Lawesson, S. O., Marshall, J. T. B., Schroll, G. & Williams, D. H. Mass spectrometry. XII. Mass spectra of enamines. *J. Chem. Soc. B* 940–946 (1966).
22. Page, C. C., Moser, C. C., Chen, X. & Dutton, P. L. Natural engineering principles of electron tunnelling in biological oxidation-reduction. *Nature* **402,** 47–52 (1999).
23. Häfelinger, G. & Mack, H.-G. in *The Chemistry of Enamines* Ch. 1 (ed. Rappaport, Z.) 1–87 (Wiley & Sons, 1994).
24. Okada, Y., Nishimoto, A., Akaba, R. & Chiba, K. Electron-transfer-induced intermolecular [2+2] cycloaddition reactions based on the aromatic "redox tag" strategy. *J. Org. Chem.* **76,** 3470–3476 (2011).
25. Tzirakis, M. D., Lykakis, I. N. & Orfanopoulos, M. Decatungstate as an efficient photocatalyst in organic chemistry. *Chem. Soc. Rev.* **38,** 2609–2621 (2009).
26. Ravelli, D., Albini, A. & Fagnoni, M. Smooth photocatalytic preparation of 2-substituted 1,3-benzodioxoles. *Chem. Eur. J.* **17,** 572–579 (2011).
27. Prudhomme, D. R., Wang, Z. & Rizzo, C. J. An improved photosensitizer for the photoinduced electron-transfer deoxygenation of benzoates and *m*-(trifluoromethyl)benzoates. *J. Org. Chem.* **62,** 8257–8260 (1997).
28. Blouin, N. & Leclerc, M. Poly(2,7-carbazole)s: structure-property relationships. *Acc. Chem. Res.* **41,** 1110–1119 (2008).
29. Yamase, T., Takabayashi, N. & Kaji, M. Solution photochemistry of tetrakis(tetrabutylammonium) decatungstate(VI) and catalytic hydrogen evolution from alcohols. *J. Chem. Soc. Dalton Trans.* 793–799 (1984).
30. Bertrand, S., Hoffmann, N. & Pete, J.-P. Highly efficient and stereoselective radical addition of tertiary amines to electron-deficient alkenes — application to the enantioselective synthesis of Necine bases. *Eur. J. Org. Chem.* 2227–2238 (2000).

Supplementary Information is available in the online version of the paper.

Acknowledgements Financial support was provided by the ICIQ Foundation, MINECO (project CTQ2013-45938-P and Severo Ochoa Excellence Accreditation 2014-2018, SEV-2013-0319), AGAUR (2014 SGR 1059), and the European Research Council (ERC 278541, ORGA-NAUT). J.J.M. and S.P. thank the Marie Curie COFUND (291787-ICIQ-IPMP) and the CELLEX Foundation, respectively, for postdoctoral fellowships. We thank M. Minozzi, M. Nappi and E. Raluy for preliminary investigations, D. Merli and D. Dondi for assistance with EPR experiments, and D. Ravelli for discussions.

Author Contributions J.J.M., D.B. and S.P. performed the experiments and analysed the data. J.J.M., D.B., S.P., and P.M. designed the experiments. M.F. and P.M. conceived the project. P.M. directed the project, and P.M. and J.J.M. wrote the manuscript with contributions from all the authors.

Author Information Crystallographic data for the iminium ion **A-1** and for compounds **3i** and **7h** have been deposited with the Cambridge Crystallographic Data Centre, accession numbers CCDC 1437991, 1437992 and 1437993, respectively. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.M. (pmelchiorre@iciq.es).

LETTER

doi:10.1038/nature17449

Chemical weathering as a mechanism for the climatic control of bedrock river incision

Brendan P. Murphy[1], Joel P. L. Johnson[1], Nicole M. Gasparini[2] & Leonard S. Sklar[3]

Feedbacks between climate, erosion and tectonics influence the rates of chemical weathering reactions[1,2], which can consume atmospheric CO_2 and modulate global climate[3,4]. However, quantitative predictions for the coupling of these feedbacks are limited because the specific mechanisms by which climate controls erosion are poorly understood. Here we show that climate-dependent chemical weathering controls the erodibility of bedrock-floored rivers across a rainfall gradient on the Big Island of Hawai'i. Field data demonstrate that the physical strength of bedrock in streambeds varies with the degree of chemical weathering, which increases systematically with local rainfall rate. We find that incorporating the quantified relationships between local rainfall and erodibility into a commonly used river incision model is necessary to predict the rates and patterns of downcutting of these rivers. In contrast to using only precipitation-dependent river discharge to explain the climatic control of bedrock river incision[5,6], the mechanism of chemical weathering can explain strong coupling between local climate and river incision.

The plausibility of strongly coupled feedbacks between climate, erosion, and tectonics over geologic timescales has most compellingly been demonstrated by numerical models[7–9]. In these and other studies, spatial differences in climate drive spatial differences in erosion rate, focusing rock uplift and ultimately affecting crustal-scale tectonic deformation of mountain ranges. A subtle but fundamental assumption of the hypothesis that climate drives rock uplift is that local climate exerts a strong control on local erosion rate. Field results are less conclusive: while many geochronological studies have found that erosion rates and local precipitation rates are empirically correlated in mountain ranges[10–12], others have found that erosion is insensitive to local climate[13]. A mechanistic understanding of how local climate actually influences landscape erosion rates is needed to explain conflicting field studies and to quantify the strength of climate–erosion feedbacks, particularly as conclusive evidence for a complete coupling of climate, erosion and tectonics remains elusive in natural landscapes[14].

River downcutting into bedrock sets landscape erosion rates in unglaciated and previously glaciated mountain ranges[15]. Without mechanistic equations quantifying the sensitivity of river incision processes to climate[16], models typically assume that climate influences local erosion rates through river discharge alone[5]. Topography can cause strong spatial precipitation gradients[17], but discharge depends on precipitation averaged over the entire upstream watershed. This results in downstream changes in discharge that are more gradual than changes in precipitation rate. Therefore, the influence of upstream-averaged precipitation rate on local discharge predicts a weaker link between climate and erosion than if local erosion rates depend on local precipitation rate[18].

We demonstrate an erosional feedback mechanism that can explain strong and local climatic control of bedrock river incision: chemical weathering can physically weaken bedrock[19], and rock strength strongly influences bedrock erodibility[20]. Increased erosion rates in turn expose fresher rock. Weathering reaction rates are dependent on the surface area of unweathered mineral grains[1] and climate, owing largely to the influence of water availability[21]. However, the implications of weathering for bedrock river erosion have only been explored at the scale of channel cross-sectional geometry[22,23], and without explicit or systematic links to climate.

The Kohala Peninsula on the Big Island of Hawai'i (Fig. 1b) is an ideal landscape in which to isolate the effects of climate and chemical weathering on fluvial incision because: (1) the bedrock is exclusively basalt with well-constrained ages and a known initial topography as it is a shield volcano[24]; (2) the landscape is tectonically quiescent and watersheds lack active internal deformation (Fig. 1c, d); (3) the base level history has been consistent for the channels we study on both sides of the peninsula, set by sea level and subsidence[25]; and (4) there is an orographic precipitation gradient with modern-day mean annual rainfall rates that span more than an order of magnitude[26] (Fig. 1a). Although absolute rainfall rates over the past 150 kyr may have been as much as 50% lower than today, the dry side of the peninsula has been consistently semiarid to arid and the direction of the orographic gradient has been constant since the formation of the volcano[21,27]. While 'climate' collectively encompasses many variables (for example, precipitation, temperature, storminess), we only consider mean annual precipitation, MAP, because it can be constrained from palaeoclimate records and influences rates of both long-term weathering[21] and fluvial incision[6]. Temperature effects on weathering across Kohala are negligible compared to precipitation, as indicated by Arrhenius equation calculations and previous work across Kohala[21,27] (Extended Data Fig. 1).

Our field analysis focuses on two channels with morphologies typical of the dry and wet sides of the peninsula[18,25] (Fig. 1a). In both dry-side Puanui Gulch and wet-side Waianaia Gulch, intermittency in streamflow often exposes bedrock in the bed and banks to the atmosphere, as is common in steep, first and second order mountain channels[28,29]. To quantify the effect of chemical weathering on erodibility, we measured rock strength in 12 river reaches (Fig. 1a) with exposed bedrock beds (in total, 542 *in situ* compressive strength measurements using a Schmidt hammer, and 117 laboratory measurements of tensile strength and bulk density from drilled rock cores). In addition, we measured rock chemistry for 45 cores collected from 4 of the sampling reaches spanning the climate gradient (Methods and Extended Data Tables 1, 2 and 3).

Chemical composition, mineralogy, and bedrock density at our sites demonstrate that chemical weathering varies systematically with local precipitation rate. Over a 2 m yr^{-1} increase in local MAP there is a progressive depletion of mobile cations (Fig. 2; Extended Data Table 1), which is indicative of silicate weathering. Sulfur shows the greatest fractional mass loss (−0.5), which we interpret to represent rapid weathering of volcanic glass. The enrichment of iron and aluminium indicates the development of silicate weathering by-products. Microscope analysis of the cores confirms a systematic increase in mineral alteration and the accumulation of iron oxides and clay minerals.

[1]Department of Geological Sciences, University of Texas at Austin, Austin, Texas 78712, USA. [2]Department of Earth and Environmental Sciences, Tulane University, New Orleans, Louisiana 70118, USA. [3]Department of Earth and Climate Sciences, San Francisco State University, San Francisco, California 94132, USA.

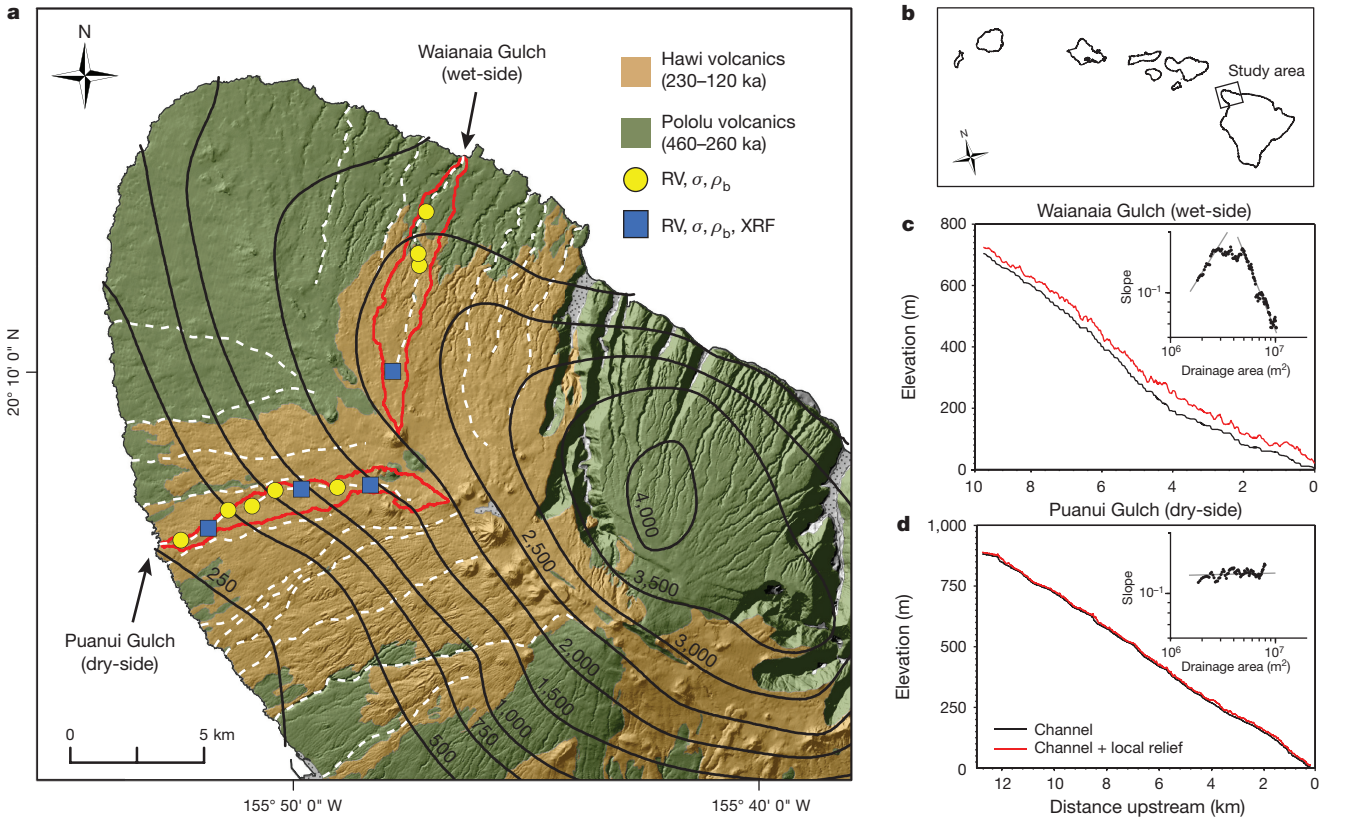


Figure 1 | Field area, measurement locations and study channels. **a**, Hillshade map including basalt units[29] (Hawi and Pololu volcanics; Methods) and mean annual precipitation (MAP) isohyets (mm yr^{-1})[27]. The two study watersheds are delineated in red, and white dashed lines indicate all 18 streams included in the stream power analysis. All study reaches (yellow circles and blue squares) have rock physical measurements (RV, σ, ρ_b); blue squares also include rock chemistry measurements from X-ray fluorescence (XRF). The locations of Waianaia Gulch and Puanui Gulch are arrowed. **b**, The study area shown within the Hawaiian Islands. **c**, Waianaia Gulch longitudinal profile (black) and the average adjacent ridge elevations (red). Inset, slope–area plot shows best-fit trends for convex and concave channel slopes[18,25]. **d**, Puanui Gulch longitudinal profile (black) and the average adjacent ridge elevation (red). Inset, slope–area plot shows the relatively uniform channel slope with best-fit trend.

Furthermore, the average dry bulk density, ρ_b, of bedrock at the surface of the cores decreases as a power-law function with increasing local MAP ($R^2 = 0.82$, $p < 0.001$; Extended Data Fig. 2; Extended Data Table 2), indicating mass loss coincident with the observations of chemical leaching and alteration.

Rock strength decreases systematically across the climate gradient with increases in chemical weathering (Extended Data Table 3). Along the dry-side Puanui Gulch, local MAP and rock strength are strongly correlated and well fitted by a power law ($R^2 = 0.90$, $p < 0.0005$; Fig. 3a). Rock strength decreases by more than 50% over nearly an order of magnitude increase in local MAP. However, wet-side Waianaia strengths do not match the dry-side trend. Rock strengths from two of the four Waianaia sites are greater than predicted by the dry-side regression, while the other two fall within the 95% confidence intervals of the regression. A key result is that rock strength varies systematically not only with weathering, but also as a function of local incision rate (Methods; Extended Data Table 4). Along wet-side Waianaia Gulch, rock strength increases by 79% as time-averaged incision rates increase from 0.05 mm yr^{-1} to 0.17 mm yr^{-1} ($R^2 = 0.98$, $p < 0.01$; Fig. 3b). Strength increases with incision rate because the degree of weathering decreases with depth below the bedrock surface and erosion exposes fresher, stronger rock beneath. However, the strongest wet-side rock is still 24% weaker than the least weathered dry-side rock. We interpret this to mean that the exhumation of fresh rock combined with high local precipitation rates sustain high weathering rates and thus weaker rock. This positive feedback is critical in explaining how higher incision rates can be maintained on the wet-side.

Climate-dependent weathering dominates the dry-side strength signal even though weathering rates are lower than on the wet-side.

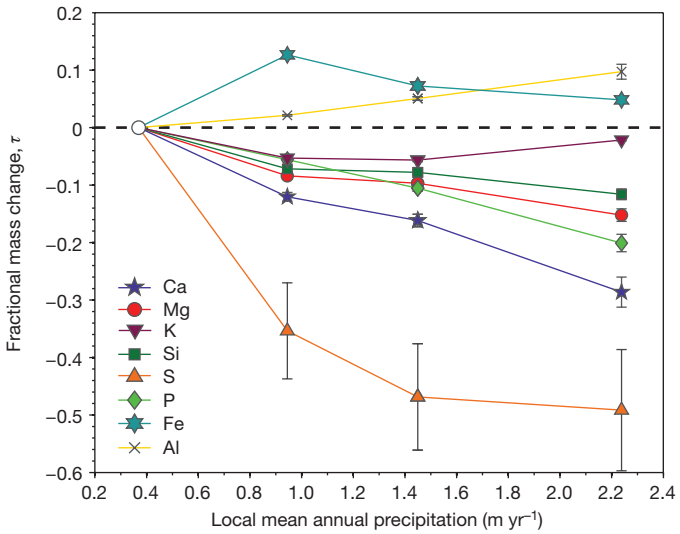


Figure 2 | Bedrock chemistry as a function of local MAP. Fractional mass changes, τ, of major elements in Hawi basalt are shown with local MAP. All points represent an average of $n \geq 13$, except site at MAP = 2.24 ($n = 4$). Error bars represent standard error; however, some are smaller than their associated symbol. The lowest MAP site (white dot) represents parent material. Consistent with silicate weathering and the development of weathering by-products, there is a relative depletion of all major labile cations with increasing MAP and an accumulation of more immobile iron and aluminium.

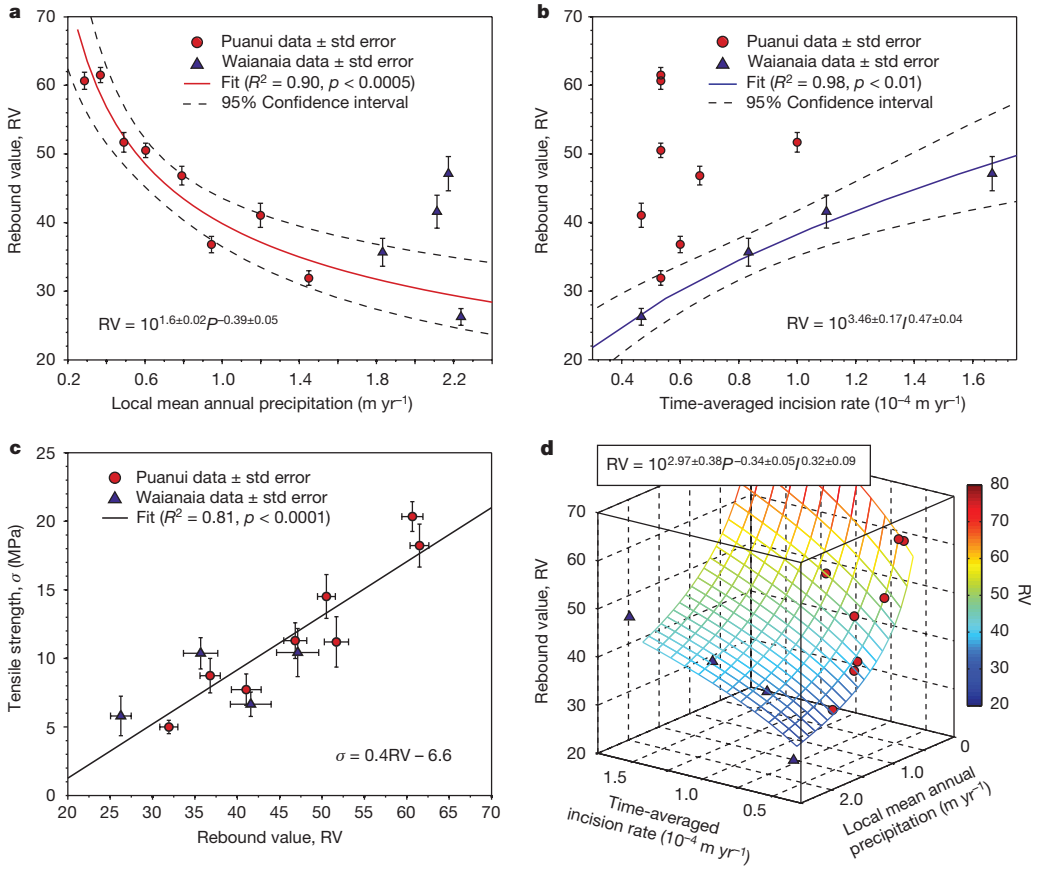


Figure 3 | Characterization of controls on rock strength. a, The average Schmidt hammer rebound values, RV, for each sampling site ($n \geq 22$) with error bars representing standard error of replicates. Puanui Gulch data are well described by a power-law regression with local MAP ($R^2 = 0.90$, $p < 0.0005$). **b**, Rebound values from Waianaia Gulch strongly follow power-law scaling with incision rate ($R^2 = 0.98$, $p < 0.01$). **c**, Average rock tensile strength, σ, for each site ($n \geq 5$) scales linearly with average rebound value ($R^2 = 0.81$, $p < 0.0001$). **d**, The best-fit surface from multiple linear regression of MAP and incision rate against average rebound values for all sites ($R^2 = 0.84$, $p < 0.0005$). Mesh colour gradient represents the rebound values predicted by the regression.

The weathering signature is well preserved along the dry-side channel because precipitation rates progressively increase towards the headwaters of the dry-side watershed so the more weathered rock is found at progressively smaller drainage areas. Incision rates on the dry-side are not only lower (an average of 0.06 mm yr^{-1}) but also fairly spatially uniform (Figs 1d, 3b), which reflects the trade-off between weaker rock upstream and lower (on average) discharges.

Combining the data from both sides of the peninsula, we use multiple linear regression and find that rock strength is well-predicted as a power function of both local precipitation and incision rate ($R^2 = 0.84$, $p < 0.0005$):

$$\mathrm{RV} = 10^{2.97\pm0.38}P^{-0.34\pm0.05}I^{0.32\pm0.09} \quad (1)$$

where RV is the Schmidt hammer rebound value, I is the time-averaged incision rate (m yr^{-1}), P is the local MAP (m yr^{-1}), and best-fit exponents are given ± standard error (Fig. 3d). The statistical significance of the regression ($p < 0.0005$) demonstrates that these 12 channel reaches are sufficient to quantify the correlation among variables, and together, incision rate and precipitation rate explain 84% of the variability in rock strength observed across the peninsula. While equation (1) is calibrated for Kohala bedrock channels alone, this conceptual framework suggests the possibility of constraining incision rates from rock strengths and local precipitation rate.

To evaluate how strength variations influence patterns of incision, rock strength must be related to bedrock erodibility. Laboratory experiments[20] show that fluvial abrasion rates, I_A, vary inversely with the square of rock tensile strength σ such that $I_A \propto \sigma^{-2.0\pm0.1}$. We find that Schmidt hammer rebound values are linearly correlated with laboratory measurements of rock tensile strength ($R^2 = 0.81$, $p < 0.0001$; Fig. 3c) (Methods). Assuming in the absence of other constraints that the combination of active incision processes (for example, abrasion and block plucking) scales like abrasion alone ($I \approx I_A$), we combine our field-based relation (equation (1)) with $I \propto \sigma^{-2.0\pm0.1}$ and RV $\propto \sigma$ to obtain:

$$I \propto P^{0.42\pm0.13} \quad (2)$$

where the uncertainty represents propagated standard error. All else being equal, equation (2) predicts how the rate of river incision into Kohala basalt should scale with local MAP owing to the effects of precipitation-driven chemical weathering on rock erodibility.

Finally, we incorporate local precipitation-dependent effects on bedrock erodibility into a river incision model and demonstrate that it greatly improves predictions of incision rates, first for the two study channels and then for all 18 major channels in our broader Kohala study area (Fig. 1a). The commonly used stream power model treats fluvial incision rate, I, as a power-law function of channel slope, S, and discharge[30]. On the much older Hawaiian island of Kauai, recent work has suggested that climate influences incision through the effect of rainfall gradients on discharge[6], represented in the stream power model below as the product of upstream-averaged mean annual precipitation rate, $\overline{P}$, and drainage area, A:

$$I = K(\overline{P}A)^m S^n \quad (3)$$

where K is a bedrock erodibility coefficient and m and n are positive exponents[30]. We modify equation (3) to explicitly incorporate the effect of local mean annual precipitation rate, P, on erodibility:

$$I = (K_i P^d)(\overline{P}A)^m S^n \quad (4)$$

where K_i is the precipitation-independent component of erodibility. Equation (2) suggests that precipitation exponent d should equal 0.42. Multiple linear regression was conducted to find the best-fit parameters (d, K_i, m, n) for relating modelled to digital elevation model (DEM)-derived incision rates (Methods). Three erodibility scenarios were tested (Extended Data Table 5): $d = 0$ (equivalent to equation (3)), $d = 0.42$, and d as a free parameter. For each of these three scenarios, two regressions were calculated. First, exponents m and n were set to 0.5 and 1 respectively; these values are commonly used for eroding fluvial systems[6,30]. Second, m and n were treated as free parameters. Best-fit K_i was calculated in all cases.

By including $d = 0.42$ from the rock strength analysis, equation (4) can explain 69% of the variability in incision rates for Waianaia and Puanui (Extended Data Table 5). The fit only improves slightly when d is a free parameter ($R^2 = 0.75$). In contrast, the conventional stream power model ($d = 0$) cannot provide acceptable fits to the study channels ($R^2 = 0.19$). Our modification to the stream power model (equation (4)) does not change the nature of erosion rate dependence on slope and drainage area, as best-fit m and n exponents in regressions that include precipitation-dependent erodibility ($d > 0$) remain close to the commonly used values ($m = 0.47$ and 0.5, $n = 0.94$ and 1.15; Extended Data Table 5). Best-fit values for precipitation exponent d are 0.75 and 0.76, suggesting that the influence of local MAP on the channel incision rate in these channels is even greater than predicted by our rock strength analysis (equation (2)). Channel incision may be more sensitive to weathering than predicted from the abrasion-strength scaling assumption, because perhaps weathering also enhances block plucking or coarse sediment supply from local hillslopes, plausibly enhancing incision rates[20].

To test whether climate-dependent weathering can explain incision patterns across the Kohala Peninsula, we expand the analysis to include a total of 18 channels, including Puanui and Waianaia (Fig. 1a). Repeating the free parameter approach, we find an average best-fit value of 0.71 ± 0.12 for the precipitation exponent d (Extended Data Table 5). Average best-fit values for exponents m and n remain close to commonly used values (0.51 ± 0.02 and 1.17 ± 0.09, respectively). In every watershed, local precipitation-dependent erodibility is required ($d > 0$) to explain fluvial incision patterns across the Kohala Peninsula.

How broadly applicable are these results to other locations and lithologies? At our field site, flow is intermittent, which provides extended periods of time for chemical weathering to deteriorate bedrock exposed across the entire channel cross-section. However, Kohala is far from unique in this regard, as intermittent flow is common in low-order mountain watersheds, regardless of MAP[28,29]. In arid and semiarid landscapes, flow can be intermittent out to large drainage areas that encompass entire mountain ranges. Our results show that chemical weathering can influence rock erodibility across a large range of MAP, not only in wet areas, since what ultimately matters is the weathering rate relative to incision rate. Seasonality of precipitation can also expose bedrock at larger drainage areas for portions of the year. Depending on rock and water chemistry, it may also be possible for weathering to occur under water on portions of the streambed under base flow. However, base flow rarely covers the entire extent of the streambed. Even if base flow inhibits weathering in the lowest elevations of the channel cross-section (thalweg), chemical weathering processes could still enhance bedrock erodibility on the majority of the subaerially exposed but still active streambed[22,23].

This study focuses entirely on basalt, which can be relatively susceptible to weathering when fresh. In other landscapes, including tectonically active settings, lithology is more heterogeneous and the influence of this rock variability on weathering driven changes in river erodibility remains uncertain. However, chemical weathering is ubiquitous, and correlations between the degree of weathering and rock strength have been measured in a range of rock types[19,31]. Chemical weathering is therefore a plausible mechanism for the climatic control of river downcutting in other settings. Owing to variations in lithologic susceptibility to weathering, different landscapes or even different areas within a single landscape may have different sensitivities to climate. If climate-dependent chemical weathering is controlling river incision, then lithologic variability may explain conflicting observations of climate-erosion coupling found in previous research[12,13].

The range of long-term fluvial incision rates across the Kohala Peninsula (about 0.05–0.5 mm yr^{-1}) is representative of many erosional landscapes of the world, but does not reach the highest rates found in actively uplifting mountain belts (about 1–10 mm yr^{-1}). If erosion rates far exceed chemical weathering rates, then bedrock erodibility may become insensitive to climate. Consequently, river incision may be most sensitive to climate in regions where erosion rates and weathering rates are comparable. We note that the weathering and erosion of hillslopes is also dependent on climate and lithology and will control the supply of coarse sediment to channels[32], another modulator of river downcutting[20]. Therefore, sediment supply is another way in which climate-controlled chemical weathering could influence the rates and patterns of landscape erosion, and introduce variability in the sensitivity of landscapes to climate.

Chemical weathering provides a strong control on bedrock erodibility and the patterns of fluvial incision across the Kohala Peninsula. The strength of this mechanism may vary with tectonic, lithologic and climatic conditions, and thus requires further evaluation across a diverse range of landscapes. The scaling relations we present here quantify climatic controls on rock erodibility and provide a straightforward method for both assessing climatic controls from topographic data and for incorporating climatic feedbacks into landscape evolution models.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 8 July 2015; accepted 23 February 2016.

1. Riebe, C. S., Kirchner, J. W. & Finkel, R. C. Erosional and climatic effects on long-term chemical weathering rates in granitic landscapes spanning diverse climate regimes. *Earth Planet. Sci. Lett.* **224,** 547–562 (2004).
2. Dixon, J. L., Hartshorn, A. S., Heimsath, A. M., DiBiase, R. A. & Whipple, K. X. Chemical weathering response to tectonic forcing: a soils perspective from the San Gabriel Mountains, California. *Earth Planet. Sci. Lett.* **323–324,** 40–49 (2012).
3. Wallmann, K. Controls on the Cretaceous and Cenozoic evolution of seawater composition, atmospheric CO_2 and climate. *Geochim. Cosmochim. Acta* **65,** 3005–3025 (2001).
4. Gaillardet, J., Dupré, B., Louvat, P. & Allegre, C. Global silicate weathering and CO_2 consumption rates deduced from the chemistry of large rivers. *Chem. Geol.* **159,** 3–30 (1999).
5. Roe, G. H., Montgomery, D. R. & Hallet, B. Effects of orographic precipitation variations on the concavity of steady-state river profiles. *Geology* **30,** 143–146 (2002).
6. Ferrier, K. L., Huppert, K. L. & Perron, J. T. Climatic control of bedrock river incision. *Nature* **496,** 206–209 (2013).
7. Koons, P. O. The topographic evolution of collisional mountain belts: a numerical look at the Southern Alps, New Zealand. *Am. J. Sci.* **289,** 1041–1069 (1989).
8. Beaumont, C., Fullsack, P. & Hamilton, J. in *Thrust Tectonics* (ed. McClay, K. R.) 1–18 (Chapman and Hall, 1992).
9. Willett, S. D. Orogeny and orography: the effects of erosion on the structure of mountain belts. *J. Geophys. Res.* **104,** 28957–28981 (1999).
10. Reiners, P. W., Ehlers, T. A., Mitchell, S. G. & Montgomery, D. R. Coupled spatial variations in precipitation and long-term erosion rates along the Washington Cascades. *Nature* **426,** 645–647 (2003).
11. Moon, S. *et al.* Climatic control of denudation in the deglaciated landscape of the Washington Cascades. *Nature Geosci.* **4,** 469–473 (2011).
12. Thiede, R. C., Bookhagen, B., Arrowsmith, J. R., Sobel, E. R. & Strecker, M. R. Climatic control on rapid exhumation along the Southern Himalayan Front. *Earth Planet. Sci. Lett.* **222,** 791–806 (2004).
13. Burbank, D. W. *et al.* Decoupling of erosion and precipitation in the Himalayas. *Nature* **426,** 652–655 (2003).

14. Whipple, K. X. The influence of climate on the tectonic evolution of mountain belts. *Nature Geosci.* **2,** 97–104 (2009).
15. Howard, A. D., Dietrich, W. E. & Seidl, M. A. Modeling fluvial erosion on regional to continental scales. *J. Geophys. Res.* **99,** 13971–13986 (1994).
16. Whipple, K. X., Hancock, G. S. & Anderson, R. S. River incision into bedrock: mechanics and relative efficacy of plucking, abrasion, and cavitation. *Geol. Soc. Am. Bull.* **112,** 490–503 (2000).
17. Galewsky, J. Rain shadow development during the growth of mountain ranges: an atmospheric dynamics perspective. *J. Geophys. Res.* **114,** F01018 (2009).
18. Han, J., Gasparini, N. M., Johnson, J. P. L. & Murphy, B. P. Modeling the influence of rainfall gradients on discharge, bedrock erodibility, and river profile evolution, with application to the Big Island, Hawai'i. *J. Geophys. Res.* **119,** 1418–1440 (2014).
19. Moon, V. & Jayawardane, J. Geomechanical and geochemical changes during early stages of weathering of Karamu Basalt, New Zealand. *Eng. Geol.* **74,** 57–72 (2004).
20. Sklar, L. S. & Dietrich, W. E. Sediment and rock strength controls on river incision into bedrock. *Geology* **29,** 1087–1090 (2001).
21. Chadwick, O. A. *et al.* The impact of climate on the biogeochemical functioning of volcanic soils. *Chem. Geol.* **202,** 195–223 (2003).
22. Montgomery, D. R. Observations on the role of lithology in strath terrace formation and bedrock channel width. *Am. J. Sci.* **304,** 454–476 (2004).
23. Small, E. E., Blom, T., Hancock, G. S., Hynek, B. M. & Wobus, C. W. Variability of rock erodibility in bedrock-floored stream channels based on abrasion mill experiments. *J. Geophys. Res. Earth Surf.* **120,** 1455–1469 (2015).
24. Wolfe, E. W. & Morris, J. *Geologic Map of the Island of Hawaii.* Map I–2524-A (US Geological Survey, 1996).
25. Menking, J. A., Han, J., Gasparini, N. M. & Johnson, J. P. L. The effects of precipitation gradients on river profile evolution on the Big Island of Hawai'i. *Geol. Soc. Am. Bull.* **125,** 594–608 (2013).
26. Giambelluca, T. W. *et al.* Online Rainfall Atlas of Hawai'i. *Bull. Am. Meteorol. Soc.* **94,** 313–316 (2013).
27. Porder, S., Hilley, G. E. & Chadwick, O. A. Chemical weathering, mass loss, and dust inputs across a climate by time matrix in the Hawaiian Islands. *Earth Planet. Sci. Lett.* **258,** 414–427 (2007).
28. Vance, L. K. *Geographically Isolated Wetlands and Intermittent/Ephemeral Streams in Montana: Extent, Distribution, and Function* http://dx.doi.org/10.5962/bhl.title.51000 (Montana Natural Heritage Program, Helena, Montana, 2009).
29. Caruso, B. S. GIS-based stream classification in a mountain watershed for jurisdictional evaluation. *J. Am. Water Resour. Assoc.* **50,** 1304–1324 (2014).
30. Whipple, K. X. & Tucker, G. E. Dynamics of the stream-power river incision model: implications for height limits of mountain ranges, landscape response timescales, and research needs. *J. Geophys. Res.* **104,** 17661–17674 (1999).
31. Basu, A., Ghosh, N. & Das, M. Categorizing weathering grades of quartzitic materials and assessing Brazilian tensile strength with reference to assigned grades. *Int. J. Rock Mech. Min. Sci.* **49,** 148–155 (2012).
32. Riebe, C. S., Sklar, L. S., Lukens, C. E. & Shuster, D. L. Climate and topography control the size and flux of sediment produced on steep mountain slopes. *Proc. Natl Acad. Sci. USA* **112,** 15574–15579 (2015).

Acknowledgements This work was supported by NSF grant EAR-1024982 to J.P.L.J., NSF grant EAR-1025055 and a Tulane Research Enhancement grant to N.M.G., and an NSF Graduate Research Fellowship to B.P.M. Airborne LiDAR was acquired by NCALM through a Seed grant to B.P.M. We thank J. Pipan for his work, H. Rowe for his XRF equipment, and landowners (Kohala Institute at 'Iole, Ponoholo Ranch, and Parker Ranch) for access, support and assistance. We also thank D. Mohrig and D. Breecker for reviews, and L. Olinde, J. Han, G. Fischer, J. Adams, I. Yokelson, and K. Kirchner for assistance in the field.

Author Contributions J.P.L.J. and N.M.G. conceived the project. B.P.M. conducted the fieldwork, laboratory work, and data analysis. L.S.S. contributed to the analysis and incorporation of rock strength data. B.P.M. wrote the manuscript with interpretations and contributions from all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.P.M. (bpmurphy@utexas.edu).

METHODS

Site selection. We selected a total of 12 channel reaches: 8 from Puanui Gulch and 4 from Waianaia Gulch. Selection was based on location along the precipitation gradient, valley relief, accessibility, and exposure of bedrock in the channel. To reduce the likelihood of user bias in the selection of individual points for data collection, measurements and samples were collected at equal intervals along 25–50 m transects. Transects were laid out with a measuring tape parallel to the flow direction on bedrock exposed in the streambed, where field evidence suggested the area was dominated by fluvial processes. In both wet- and dry-side channels we observe extensive fluting, sculpting and detached blocks in streambeds, which indicate that a combination of impact wear and hydraulic plucking cause fluvial downcutting.

Waianaia Gulch on the wet-side has incised deeply into both Hawi basalt (230–120 ka) and the older Pololu basalt (460–260 ka)[33,34], with valley depth exceeding 50 m in places. Like other wet-side channels it has a convex-concave longitudinal profile (Fig. 1c) that is not primarily controlled by the lithologic contact between Hawi and Pololu[25]. Across the wet-side watershed, mean annual precipitation (MAP) varies by less than a factor of two, from 1,500 to 2,300 mm yr^{-1} (ref. 26). Rainfall is frequent and stream flow occurs more often than on the dry side, but discharge is intermittent and only common after soils are saturated from multiple days of substantial rainfall[25]. In similar rivers on the Big Island of Hawai'i, stream water chemistry suggests that rivers are fed only by surface water and shallow groundwater[35], as groundwater tables in the Hawaiian Islands are close to sea level[36].

The dry-side Puanui Gulch has incised only into Hawi basalt. Local MAP within this watershed varies by almost an order of magnitude, from 25 mm yr^{-1} to 2,000 mm yr^{-1} (ref. 26). Only large, intense and infrequent storms generate discharge[25]. The preservation of original basalt features, such as flow lobes, outside of channels indicate that hillslope processes have minimally modified the dry-side topography. Local relief demonstrates that dry-side channels have incised <10 m in most places. Because fluvial incision has been minor, the channel slopes do not exhibit notable variation with drainage area (Fig. 1d) and are still consistent with the initial slope of the shield volcano.

Knickpoints—most 1 to 3 m high—occur on relatively thicker basalt flow beds composed of large fracture-bounded blocks[25]. Much larger profile-scale knickpoints (often initiated by landslides) occur elsewhere in Hawai'i[37–39] but not along the channels we study (Fig. 1c, d). Although some areas on the east coast of Kohala have small sea cliffs ($\sim$10 m relief) possibly created by wave erosion, all of the channels that we consider have fully cut down through these sea cliffs. The streams are therefore not connected to any effect of wave erosion, and maintain a base level that is connected to sea level. This is confirmed by field observations of the study channels. We also intentionally avoid streams in the region of the Pololu Slump, as base level of these channels was affected by a massive landslide and some are still disconnected from the ocean by large sea cliffs (400–500 m relief)[37].

Schmidt hammer. The type-N Schmidt hammer is a handheld tool that measures the *in situ* elastic properties of the bedrock and provides a measurement that scales with rock compressive strength[40]. The Schmidt hammer measures the percentage distance of rebound of a spring-loaded mass after it impacts a surface. This measurement is referred to as the rebound value, RV. According to the manufacturer, Schmidt hammer readings between 20 RV and 55 RV correspond to compressive strengths of 10–70 MPa.

Data were collected at equal distance intervals along transects. Intervals were typically between 0.5 to 1 m, with the exact distance chosen to target a total of 50 measurements along the length of exposed bedrock. This sample size exceeds the minimum statistical requirement ($n = 33$) suggested for Schmidt hammer measurements on basalt[41]. However, at some wet-side sampling sites, bedrock exposure or rock quality limited the number of possible measurements. Exact measurement location was only adjusted if the sampling point fell on or near an edge or major fracture, in which case the measurement was moved along transect to a distance of 10 cm from the edge or fracture, consistent with instrument use recommendations[42].

Tensile strength. At every site approximately 10–20 bedrock cores were drilled. The exact number of samples depended on how much gasoline and water we were able to carry to power and cool the rock drill. Each core was 2.54 cm in diameter and varied in depth from 10 cm to 20 cm. Cores were extracted perpendicular to the rock surface at equal intervals along transect. The distance between cores was typically between 1.5 m and 3 m, but was chosen to be a multiple of the Schmidt hammer measurement distance. This was done for consistency in comparing rock property measurements. The exact drilling locations were directly adjacent to but never on a Schmidt hammer measurement location, in order to reduce any potential effect of impact damage by the Schmidt hammer on tensile strength measurements.

The Brazilian splitting test was used to measure the tensile strength of cores from all 12 sites. This laboratory method requires the measurements of cylinders with a diameter-to-depth ratio less than 2, so cores were cut perpendicular to the long axis at 2.54 cm intervals to create cylinders with a 1:1 ratio[43]. In some cases, the *in situ* competence of the rock was so deteriorated (mainly in highly weathered cores from Waianaia Gulch) that they broke when drilled in the field. These cores were not used for measurement, reducing the sample size for those sites. All of the cut segments of the cores were first dried (see Methods section 'Bulk density') to reduce any potential effect of moisture on strength measurements. The tensile strengths that we present only represent measurements made on the uppermost segments of the cores, or the 2.54 cm nearest the surface of the rock. After each test the geometry of the fracture was recorded and characterized. If the principal mode of failure was not a centrally located fracture parallel to the plane of loading stress[44], then the result was characterized as being of poor quality and was excluded from our analysis. The reported average tensile strengths only represent measurements of acceptable quality, which further reduced sample sizes at some sites.

Bulk density. After the cores were cut into segments but before the Brazilian splitting test, the dry bulk density, ρ_b, was measured for each segment. First, core segments were oven dried for 24 h at 105 °C, and then the dimensions and mass of each was measured with callipers and a digital scale. As with the tensile strengths, bulk density data represent the average for bedrock from the surface to a depth of 2.54 cm.

Rock chemistry. The abundance of major and trace elements were measured using energy-dispersive X-ray fluorescence (ED-XRF)[45] on the same cores used for tensile strength testing. Owing to limited machine time, we only measured rock chemistry on cores from four of the sampling sites. The subset of sites was chosen to reasonably span the range of MAP within the Hawi basalt. We had no unweathered samples of Pololu basalt as a baseline for calculations of fractional mass loss, so ED-XRF analysis was only conducted on Hawi basalts.

ED-XRF data were measured on the same bedrock cores used in the tensile strength analysis. Measurements were made on the upper face of the core, or the surface of the rock. The measurements from each sampling site were averaged to reduce any possible signals due to mineralogical heterogeneity. Comparing fractional mass loss between sites required the determination of a representative parent material. Given the demonstrated effect of local precipitation rate on soil weathering rates across Kohala[21,27,46], we assume the lowest precipitation site in Puanui is the most representative of parent material for the Hawi basalt. However to justify this assumption, we compared the average dry bulk density of that site (2.63 g cm^{-3}) to published values for fresh Hawi basalt (2.6–2.7 g cm^{-3})[33]. Additionally, XRF data demonstrate this site has the greatest abundance of major labile elements. It may be noted that sodium, a major labile cation, is missing from our analysis, but this is because it is not accurately measured with ED-XRF.

To determine the extent of fractional mass losses or gains in the basalt, we calculated a mass transfer coefficient, τ, based on the concentration of niobium, Nb, which has been shown to be the most immobile trace element in the Kohala landscape[47]:

$$\tau = \left(\frac{i_w \mathrm{Nb_p}}{i_p \mathrm{Nb_w}} - 1 \right) \quad (5)$$

where i_w and i_p are the concentrations of element i in the weathered sample and parent material, respectively, and $\mathrm{Nb_w}$ and $\mathrm{Nb_p}$ are the corresponding concentrations of niobium[2].

Microscope analysis. Thin sections were ground from the internal surfaces of the basalt cores broken during the Brazilian splitting test. Cores were impregnated with epoxy before grinding in order to preserve the morphology of delicate weathered vesicles and the secondary clay minerals that could partially fill them in more weathered samples.

Time-averaged incision rates. Kohala streams exhibit a parallel drainage pattern with relatively narrow watersheds and minimally eroded interfluves as indicated by preserved basalt flow morphologies in many places. Therefore, valley relief, or the vertical distance between present day stream elevations and adjacent interfluve or ridge elevations, could be used to estimate the depth of cumulative incision without needing to interpolate surfaces across valleys. Time-averaged incision rates were then quantified following previous methods for Hawaiian shield volcano erosion[6]. The ages of the basalts used were based on previous K–Ar dating[21,33,34]. For regions in the Pololu basalt we used an average age of 300 kyr. For Hawi basalt, we used an average age of 150 kyr, consistent with previous work[21,27,46].

Local valley relief was calculated using a moving window in ArcGIS and the best available topographic data sets for each watershed. Owing to systematic differences in valley width and depth on the two sides of the peninsula, we could not use a constant moving window size; a window big enough to measure the depth of the deeper valleys of the wet-side would erroneously report the relief within that window due to the overall slope of the shield volcano rather than valley relief

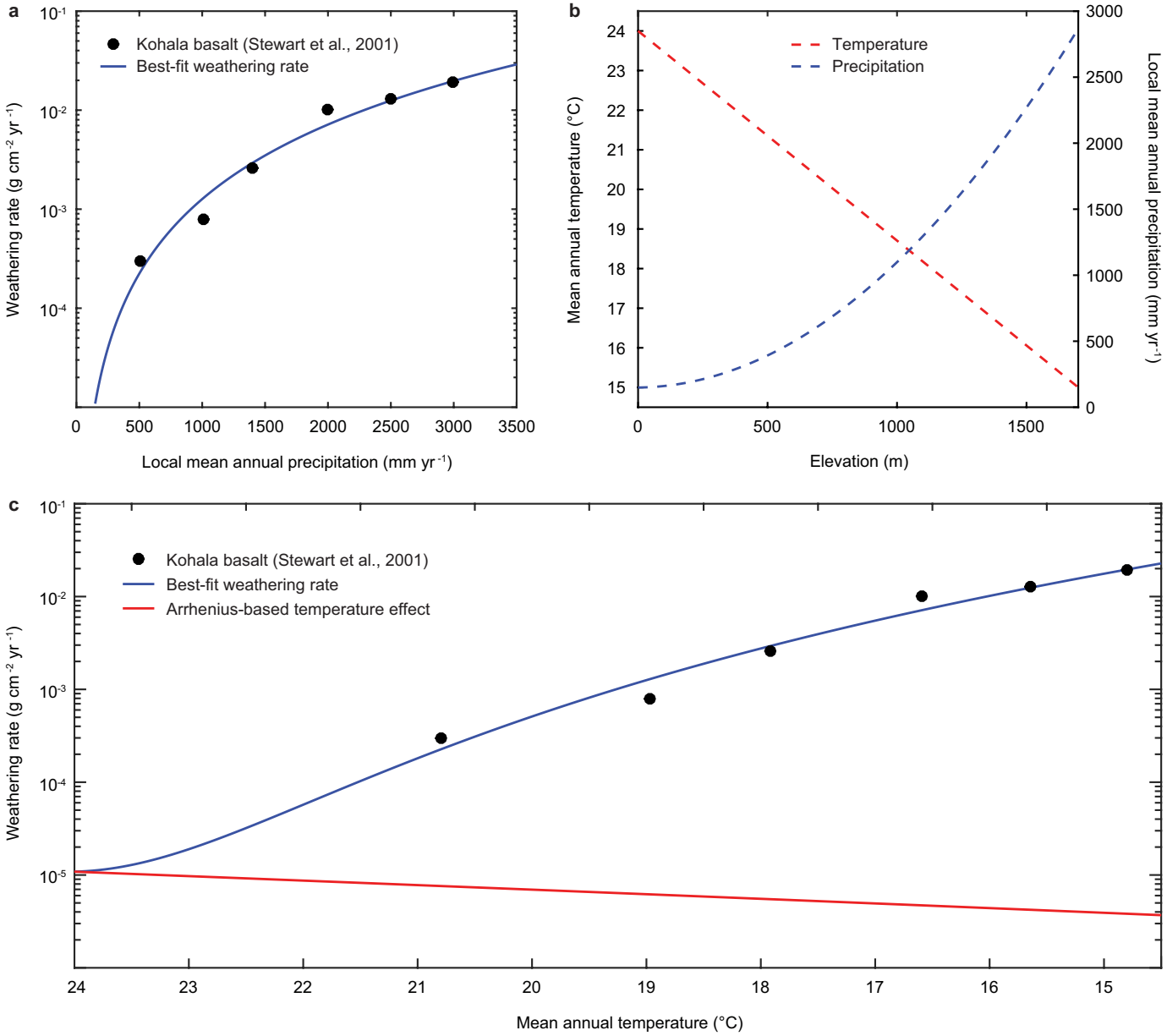
on the dry-side. Therefore, across Puanui Gulch we used a 40 m × 40 m moving window over a NCALM airborne-LiDAR derived DEM with 1-m resolution. This topography data is open source and can be downloaded at http://opentopo.sdsc.edu/datasetMetadata? otCollectionID = OT.012014.26904.1. For the remainder of the dry-side we used the same size moving window but used 1/3-arcsecond (~10 m) USGS DEMs. Across Waianaia Gulch and the rest of the wet-side we used a 200 m × 200 m moving window over a 1/3-arcsecond USGS DEM. This topography data is available for download at http://viewer.nationalmap.gov/basic/.

Regressions. Multiple linear regressions for the various forms of the stream power model were first conducted on data derived from GIS analysis along the entirety of the two study channels (that is, not just at the 12 discrete sampling reaches). Present day channel slope and drainage area were extracted from the highest resolution topographic data sets available. Local MAP was extracted from GIS data published in the Hawai'i Rainfall Atlas[27]. Incision rates were extracted as described above. Before regressions were conducted, data were smoothed using a simple moving average to reduce noise in the data, particularly related to extremely localized high slopes in the LiDAR data set.

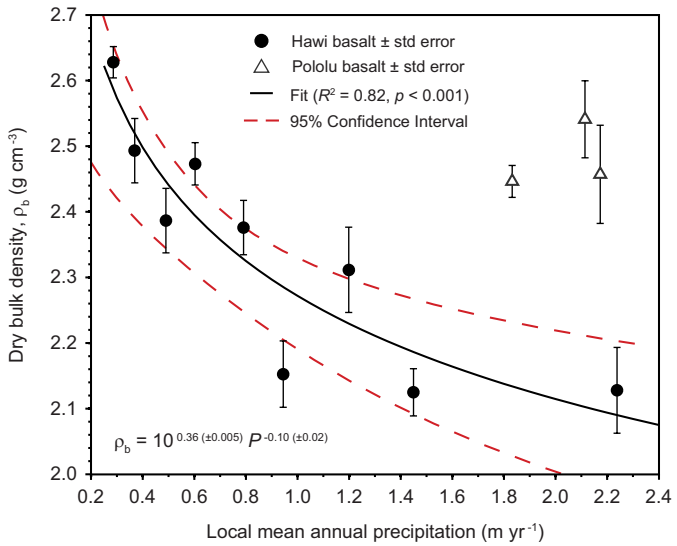
The same data were then extracted for an additional 16 channels across Kohala (using USGS DEMs for topography) and multiple linear regression was conducted on all 18 channels individually. First we used a fully free-parameter approach to get best fits for the modified stream power model. The precipitation independent erodibility coefficient, K_i, did not vary greatly between channels, which is consistent with the assumption that this value should be reasonably uniform across a landscape of uniform lithology. All 18 regressions were then repeated holding K_i constant to the average value determined from the previous set of regressions. We report the averages with standard error for best-fit exponents m, n and d when K_i is held constant (Extended Data Table 5, regressions 7–24).

Sample size. No statistical methods were used to predetermine sample size.

33. Spengler, S. R. & Garcia, M. O. Geochemistry of the Hawi lavas, Kohala Volcano, Hawaii. *Contrib. Mineral. Petrol.* **99,** 90–104 (1988).
34. McDougall, I. & Swanson, D. A. Potassium-argon ages of lavas from the Hawi and Pololu volcanic series, Kohala Volcano, Hawaii. *Geol. Soc. Am. Bull.* **83,** 3731–3738 (1972).
35. Schopka, H. H. & Derry, L. A. Chemical weathering fluxes from volcanic islands and the importance of groundwater: the Hawaiian example. *Earth Planet. Sci. Lett.* **339–340,** 67–78 (2012).
36. Oki, D. S. *Geohydrology and Numerical Simulation of the Ground-Water Flow System of Molokai, Hawaii*. Water-Resources Investigations Report 97–4176 (US Geological Survey, 1997).
37. Lamb, M. P., Howard, D., Dietrich, W. E. & Perron, J. T. Formation of amphitheater-headed valleys by waterfall erosion after large-scale slumping on Hawai'i. *Geol. Soc. Am. Bull.* **119,** 805–822 (2007).
38. Seidl, M., Dietrich, W. & Kirchner, J. Longitudinal profile development into bedrock: an analysis of Hawaiian channels. *J. Geol.* **102,** 457–474 (1994).
39. Mackey, B. H., Scheingross, J. S., Lamb, M. P. & Farley, K. A. Knickpoint formation, rapid propagation, and landscape response following coastal cliff retreat at the last interglacial sea-level highstand: Kaua'i, Hawai'i. *Geol. Soc. Am. Bull.* **126,** 925–942 (2014).
40. Aydin, A. & Basu, A. The Schmidt hammer in rock material characterization. *Eng. Geol.* **81,** 1–14 (2005).
41. Niedzielski, T., Migoń, P. & Placek, A. A minimum sample size required from Schmidt hammer measurements. *Earth Surf. Process. Landf.* **34,** 1713–1725 (2009).
42. Day, M. J. & Goudie, A. S. Field assessment of rock hardness using the Schmidt test hammer. *Br. Geomorphol. Res. Group Tech. Bull*. **18,** 19–29 (1980).
43. International Society for Rock Mechanics, Commission on Standardization of laboratory and Field Tests. Suggested methods for determining tensile strength of rock materials. *Int. J. Rock Mech. Min. Sci. Geomech. Abstr*. **15,** 99–103 (1978).
44. Rocco, C., Guinea, G. V., Planas, J. & Elices, M. Mechanisms of rupture in splitting tests. *ACI Mater. J.* **96,** 52–60 (1999).
45. Rowe, H., Hughes, N. & Robinson, K. The quantification and application of handheld energy-dispersive x-ray fluorescence (ED-XRF) in mudrock chemostratigraphy and geochemistry. *Chem. Geol.* **324–325,** 122–131 (2012).
46. Stewart, B. W., Capo, R. C. & Chadwick, O. A. Effects of rainfall on weathering rate, base cation provenance, and Sr isotope composition of Hawaiian soils. *Geochim. Cosmochim. Acta* **65,** 1087–1099 (2001).
47. Kurtz, A. C., Derry, L. A., Chadwick, O. A. & Alfano, M. J. Refractory element mobility in volcanic soils. *Geology* **28,** 683–686 (2000).



Extended Data Figure 1 | Evaluation of temperature effects on Kohala weathering rates. a, Fresh basalt weathering rates derived from measurements of modern soil weathering rates across the Kohala Peninsula[46] (filled black circles) and a best-fit power law regression to the data as a function of MAP (blue line; $R^2 = 0.96$). **b**, Variation of mean annual temperature, MAT (red dashed line), and MAP (blue dashed line) with elevation across the leeward side of the Kohala Peninsula[21]. Using the reported MAT from the coast and highest elevations of Kohala, we calculate an environmental lapse rate of 5.3 °C km^{-1}. **c**, Using the MAP and MAT relations in **b**, the measured weathering rates (filled black circles) and the best-fit relation (blue line) are replotted as a function of MAT. Using the best-fit estimate at 24 °C as a reference rate, the effect of temperature on the weathering rate was then estimated using the Arrhenius equation (red line). A threefold decrease in weathering rate is expected based on temperature alone, however measured weathering rates increase over three orders of magnitude (blue line). The measured weathering rates (filled black circles) should integrate any possible temperature effects, yet the trends are discordant, demonstrating that temperature is not a first-order control on chemical weathering rates in Kohala.



Extended Data Figure 2 | Variation of average dry bulk density with local MAP. Filled black circles represent data from sites of Hawi basalt, and open black triangles represent data from sites of Pololu basalt, with error bars showing standard error. The plotted regression is for sites in Hawi basalt ($R^2 = 0.82$, $p < 0.001$), in which bulk density decreases 20% with a 2-m increase in local MAP.

Extended Data Table 1 | Average fractional mass loss, τ and MAP

Site ID	Basalt Unit	MAP (mm yr^{-1})	n	Average τ ± std error	
				Ca	Mg
P2	Hawi	369.3	14	0	0
P6	Hawi	944.5	13	-0.120 ± 0.008	-0.084 ± 0.005
P8	Hawi	1449.3	14	-0.162 ± 0.011	-0.097 ± 0.005
W1	Hawi	2238.1	4	-0.286 ± 0.026	-0.152 ± 0.011
				K	Si
P2	Hawi	369.3	14	0	0
P6	Hawi	944.5	13	-0.053 ± 0.003	-0.072 ± 0.004
P8	Hawi	1449.3	14	-0.056 ± 0.003	-0.078 ± 0.004
W1	Hawi	2238.1	4	-0.022 ± 0.002	-0.116 ± 0.009
				S	P
P2	Hawi	369.3	14	0	0
P6	Hawi	944.5	13	-0.354 ± 0.084	-0.056 ± 0.004
P8	Hawi	1449.3	14	-0.469 ± 0.093	-0.105 ± 0.005
W1	Hawi	2238.1	4	-0.492 ± 0.105	-0.201 ± 0.015
				Fe	Al
P2	Hawi	369.3	14	0	0
P6	Hawi	944.5	13	0.127 ± 0.007	0.021 ± 0.001
P8	Hawi	1449.3	14	0.072 ± 0.004	0.051 ± 0.003
W1	Hawi	2238.1	4	0.048 ± 0.005	0.097 ± 0.013

Extended Data Table 2 | Dry bulk density, MAP and incision rate

Site ID	MAP (mm yr^{-1})	Time-Averaged Incision Rate (m yr^{-1})	Bulk Density		
			Mean (g cm^{-3})	n	Std Error (g cm^{-3})
P1	286.3	5.33E-05	2.6	10	0.02
P2	369.3	5.33E-05	2.5	10	0.05
P3	490.2	1.00E-04	2.4	11	0.05
P4	603.2	5.33E-05	2.5	11	0.03
P5	790.9	6.67E-05	2.4	13	0.04
P6	944.5	6.00E-05	2.2	11	0.05
P7	1198.4	4.67E-05	2.3	8	0.07
P8	1449.3	5.33E-05	2.1	14	0.04
W1	2238.1	4.67E-05	2.1	5	0.07
W2	1831.8	8.33E-05	2.4	7	0.02
W3	2114.0	1.10E-04	2.5	11	0.06
W4	2173.0	1.67E-04	2.5	6	0.07

Extended Data Table 3 | Rock mechanical properties, MAP and incision rate

Site ID	MAP (mm yr^{-1})	Time-Averaged Incision Rate (m yr^{-1})	Schmidt Hammer Rebound Value, RV			Tensile Strength, σ		
			Mean	n	Std Error	Mean (MPa)	n	Std Error (MPa)
P1	286.3	5.33E-05	60.7	50	1.2	20.3	10	1.1
P2	369.3	5.33E-05	61.5	50	1.1	18.2	10	1.6
P3	490.2	1.00E-04	51.7	50	1.4	11.2	11	1.8
P4	603.2	5.33E-05	50.5	50	1.1	14.5	11	1.6
P5	790.9	6.67E-05	46.8	50	1.4	11.3	13	1.3
P6	944.5	6.00E-05	36.8	50	1.2	8.7	11	1.3
P7	1198.4	4.67E-05	41.1	50	1.8	7.7	8	1.1
P8	1449.3	5.33E-05	31.9	53	1.1	5.0	14	0.5
W1	2238.1	4.67E-05	26.3	53	1.2	5.8	5	1.4
W2	1831.8	8.33E-05	35.7	40	2.0	10.4	7	1.1
W3	2114.0	1.10E-04	41.6	22	2.4	6.7	11	0.9
W4	2173.0	1.67E-04	47.1	24	2.5	10.4	6	1.8

Extended Data Table 4 | Time-averaged incision rates across the 12 field sites

Site ID	Channel Name	Basalt Unit	Age of Unit (kya)	Valley Depth (m)	Time-Averaged Incision Rate (m yr^{-1})
P1	Puanui	Hawi	150	8	5.33E-05
P2	Puanui	Hawi	150	8	5.33E-05
P3	Puanui	Hawi	150	15	1.00E-04
P4	Puanui	Hawi	150	8	5.33E-05
P5	Puanui	Hawi	150	10	6.67E-05
P6	Puanui	Hawi	150	9	6.00E-05
P7	Puanui	Hawi	150	7	4.67E-05
P8	Puanui	Hawi	150	8	5.33E-05
W1	Waianaia	Hawi	150	7	4.67E-05
W2	Waianaia	Pololu	300	25	8.33E-05
W3	Waianaia	Pololu	300	33	1.10E-04
W4	Waianaia	Pololu	300	50	1.67E-04

Extended Data Table 5 | Multiple linear regressions of stream longitudinal profiles

Regression	d	K_i	m	n	R^2	Normalized RMSE
1	0.00	2.97E-07	0.50	1.00	0.19	0.28
2	0.00	4.58E-07	0.43	0.67	0.19	0.28
3	0.42	2.78E-07	0.50	1.00	0.68	0.19
4	0.42	4.05E-07	0.47	0.94	0.69	0.19
5	0.75	2.64E-07	0.50	1.00	0.75	0.16
6	0.76	3.68E-07	0.50	1.15	0.75	0.16
7-24	0.71 ± 0.12	6.37E-07	0.51 ± 0.02	1.17 ± 0.09	0.71 ± 0.05	0.19 ± 0.03

Regressions 1–6 test variations of equation (4) with a combined data set that includes data from both Puanui and Waianaia Gulch. Regressions 7–24 show the average best-fit values with standard error for individual data regressions of 18 channels across the Kohala Peninsula. Fixed parameters are coloured red and free parameters blue. The goodness of fit is indicated by the coefficient of determination (R^2) and normalized root mean square error (RMSE). All regressions are statistically significant ($p < 0.0001$).

LETTER

doi:10.1038/nature17159

Ritual human sacrifice promoted and sustained the evolution of stratified societies

Joseph Watts[1], Oliver Sheehan[1,2], Quentin D. Atkinson[1,2], Joseph Bulbulia[3] & Russell D. Gray[1,2,4,5]

Evidence for human sacrifice is found throughout the archaeological record of early civilizations[1], the ethnographic records of indigenous world cultures[2–5], and the texts of the most prolific contemporary religions[6]. According to the social control hypothesis[2,7,8], human sacrifice legitimizes political authority and social class systems, functioning to stabilize such social stratification. Support for the social control hypothesis is largely limited to historical anecdotes of human sacrifice[2,8], where the causal claims have not been subject to rigorous quantitative cross-cultural tests. Here we test the social control hypothesis by applying Bayesian phylogenetic methods to a geographically and socially diverse sample of 93 traditional Austronesian cultures. We find strong support for models in which human sacrifice stabilizes social stratification once stratification has arisen, and promotes a shift to strictly inherited class systems. Whilst evolutionary theories of religion have focused on the functionality of prosocial and moral beliefs[9,10], our results reveal a darker link between religion and the evolution of modern hierarchical societies[11,12].

Human sacrifice—the deliberate and ritualized killing of an individual in order to please or placate supernatural beings—is known to have occurred in early Germanic, Arab, Turkic, Inuit, American, Austronesian, African, Chinese and Japanese cultures[1]. Speculation about the potential functionality of human sacrifice dates back to at least the beginning of the European colonization of Central America 500 years ago[5], and has been the subject of enduring debate across the humanities[2,13,14], social sciences[1,8,15,16] and biological sciences[17,18] ever since. The practice has been conjectured to act as a form of social catharsis[13], a justification for political conflicts[15], and, when combined with cannibalism, a means of overcoming protein shortages[16]. Political theorists have long argued that effective political authority in class-stratified societies requires legitimizing mechanisms[12,19], an idea which evolutionary scholars have recently endorsed[11,20]. According to the social control hypothesis, human sacrifice legitimizes class-based power distinctions by combining displays of ultimate authority—the taking of a life—with supernatural justifications that sanctify authority as divinely ordained[2,8,13]. Social stratification is thought to have been one of the earliest forms of institutionalized leadership to emerge in human cultures, giving rise to kingdoms, monarchies and modern political states[20,21]. Existing support for the social control hypothesis is based on anecdotal descriptions of cultures[2,8,15], and one quantitative cross-cultural study that found an association between human sacrifice and measures of social and political complexity[7]. However, this study used a sample that contained just seven cultures that practiced human sacrifice, did not control for the non-independence of cultures[7,22], and was unable to infer the direction of causality between human sacrifice and social stratification[23].

Here we test the social control hypothesis with a Bayesian phylogenetic analysis of 93 traditional Austronesian cultures from the Pulotu database[24]. Phylogenetic methods enable us to account for the common ancestry of cultures[24], test for patterns of coevolution[10,25], and infer the direction of causality based on the order that traits evolve in[23]. Austronesian cultures have been described as a natural laboratory for cross-cultural research due to the diversity of environments they inhabit and cultural features they have evolved[26]. They inhabit environments ranging from tiny atolls to continents[24], and their social structures ranged from small egalitarian, kin-based societies such as the Dobuans[24], to large, complex polities such the Hawaiians[27]. From their ancestral homeland in Taiwan, Austronesian cultures spread west to Madagascar, east to Rapa Nui, and south to New Zealand—a region covering over half the world's longitude and one-third of its latitude[24]. Their religious beliefs and practices were remarkably diverse[3,4,10], and the practice of human sacrifice was widespread throughout traditional Austronesian cultures. Common occasions for human sacrifice in these societies included the breach of taboo or custom[4], the funeral of an important chief[27], and the consecration of a newly built house or boat[3]. Ethnographic descriptions highlight that the sacrificial victims were typically of low social status, such as slaves, and the instigators were of high social status, such as priests and chiefs[3,4,27]. The methods of sacrifice included burning, drowning, strangulation, bludgeoning, burial, being crushed under a newly built canoe, being cut to pieces, as well as being rolled off the roof of a house and then decapitated[3,4,27].

For each culture in our sample, we recorded the presence or absence of human sacrifice, and coded the level of social stratification. Cultures that lacked inherited differences in wealth and status were defined as lacking social stratification, and were coded as egalitarian. Cultures were coded as moderately stratified if there were inherited differences in wealth and social position with the potential for status change within a generation, and highly stratified if there were inherited difference in wealth and social position with little or no possibility of status change within a generation (further details are provided in the Methods section). The social control hypothesis predicts that human sacrifice (i) co-evolves with social stratification, (ii) increases the chance of a culture gaining social stratification, and (iii) reduces the chance of a culture losing social stratification once stratification has arisen. Though the social control hypothesis could potentially apply to stratified societies in general[8], the hypothesis is based on descriptions of human sacrifice in highly stratified societies such as the Aztecs[2]. Here we perform two series of analyses, the first to test the effects of human sacrifice on the evolution of social stratification in general, and the second to test the effects of human sacrifice on the evolution of high social stratification (Fig. 1c, e).

We found that the extent of social stratification, as well as the presence of human sacrifice, varied throughout a wide range of geographic regions and cultural groups (Fig. 2 and Extended Data Fig. 1). Evidence of human sacrifice was observed in 40 of the 93 cultures sampled (43%). Human sacrifice was practiced in 5 of the 20 egalitarian societies (25%), 17 of the 46 moderately stratified societies (37%), and 18 of the 27 highly stratified societies (67%) sampled.

[1]School of Psychology, University of Auckland, Auckland 1142, New Zealand. [2]Max Planck Institute for the Science of Human History, Jena 07743, Germany. [3]School of Art History, Classics and Religious Studies, Victoria University of Wellington, Wellington 6014, New Zealand. [4]Research School of the Social Sciences, Australian National University, Canberra 2601, Australia. [5]Allan Wilson Centre for Molecular Ecology and Evolution, Palmerston North 4442, New Zealand.

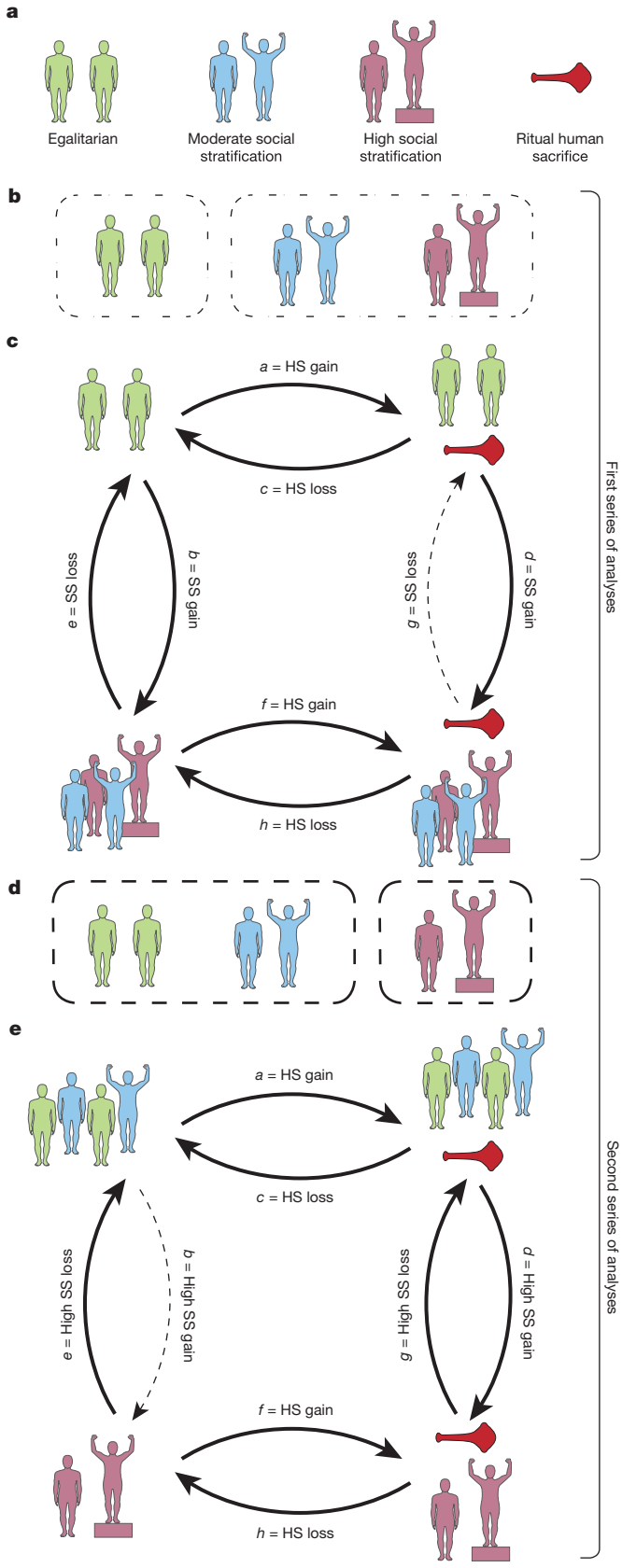


Figure 1 | Summary of the two series of analyses performed in this study. a, Key of the images used to represent social stratification and human sacrifice. **b**, In the first series of analyses, moderately and highly stratified societies cultures were grouped together to test for the co-evolution of human sacrifice with social stratification in general. **c**, Unconstrained dependent model of the co-evolution of human sacrifice (HS) and social stratification (SS) in general. The thicknesses of the arrows are proportional to the rates of change between states. **d**, In the second series of analyses, egalitarian and moderately stratified societies were grouped together to specifically test for the co-evolution of human sacrifice with high social stratification. **e**, Unconstrained dependent model of the co-evolution of human sacrifice (HS) and high social stratification (high SS). The thicknesses of arrows are proportional to the rates of change between states.

In our first series of analyses, we grouped moderate and high stratification together, referred to hereafter as 'social stratification' (Fig. 1b). To test for the co-evolution of human sacrifice and social stratification, we compared the posterior distribution of models in which human sacrifice and social stratification evolve independently of one another with models in which the two traits co-evolve such that the probability of a change in one trait is dependent on the value of the other trait[23]. We found substantial support for the models in the dependent analyses, in which human sacrifice can co-evolve with social stratification, compared with the models in the independent analyses (Bayes factor (BF) = 3.78). This indicates that human sacrifice co-evolved with social stratification. We then performed two additional constrained analyses to test whether human sacrifice functioned to drive and stabilize the evolution of social stratification, as the social control hypothesis predicts. In the first constrained analysis, cultures with and without human sacrifice were forced to have an equal chance of losing social stratification (rates *e* and *g* in Fig. 1c were set to be equal). The resulting models fitted substantially more poorly than the unconstrained dependent analyses (BF = 2.30), and did not fit substantially better than the models in the independent analysis (BF = 1.48). This indicates that human sacrifice affects the rate at which cultures lose social stratification. The unconstrained dependent model shows that cultures with human sacrifice were less likely to lose social stratification than were cultures that lacked human sacrifice (in Fig. 1c rate *e* is higher than rate *g*). In the second constrained analysis, the rate at which cultures with and without human sacrifice gained social stratification was forced to be equal (rates *b* and *d* in Fig. 1c were set to be equal). The resulting models were substantially more likely than were models in the independent analysis (BF = 4.68), and slightly more likely than models in the unconstrained dependent analysis, though not substantially so (*BF* = 0.60). Together these results indicate that human sacrifice functioned to stabilize social stratification once it had arisen, but did not affect whether egalitarian cultures gained social stratification (in Fig. 1c, rate *e* is higher than rate *g*).

In our second series of analyses, we used the same approach to test whether human sacrifice co-evolves with high social stratification specifically. In this series, we grouped egalitarian and moderately stratified societies together (Fig. 1d). We found strong support for the models in the dependent analyses over the models in the independent analyses (BF = 6.04), indicating that human sacrifice has co-evolved with high social stratification. To test the prediction that human sacrifice functions to stabilize and drive high social stratification, we performed the same sequence of constrained analyses as previously described for social stratification in general. In the first constrained analysis, cultures with and without human sacrifice were forced to have an equal chance of losing high social stratification (rates *e* and *g* in Fig. 1e were equal). The resulting models were more likely than those in the independent analysis (BF = 6.96) and the unconstrained dependent analysis (BF = 0.92), though the difference was only substantial in the case of the former. This indicates that the presence of human sacrifice is not associated with a change in the rate at which highly stratified cultures become less stratified. The second analysis was constrained so that cultures with and without human sacrifice were forced to have an equal chance of gaining high social stratification (rates *b* and *d* in Fig. 1e are equal). The resulting models were substantially less likely than were the models in the unconstrained dependent analysis (BF = 4.70), and

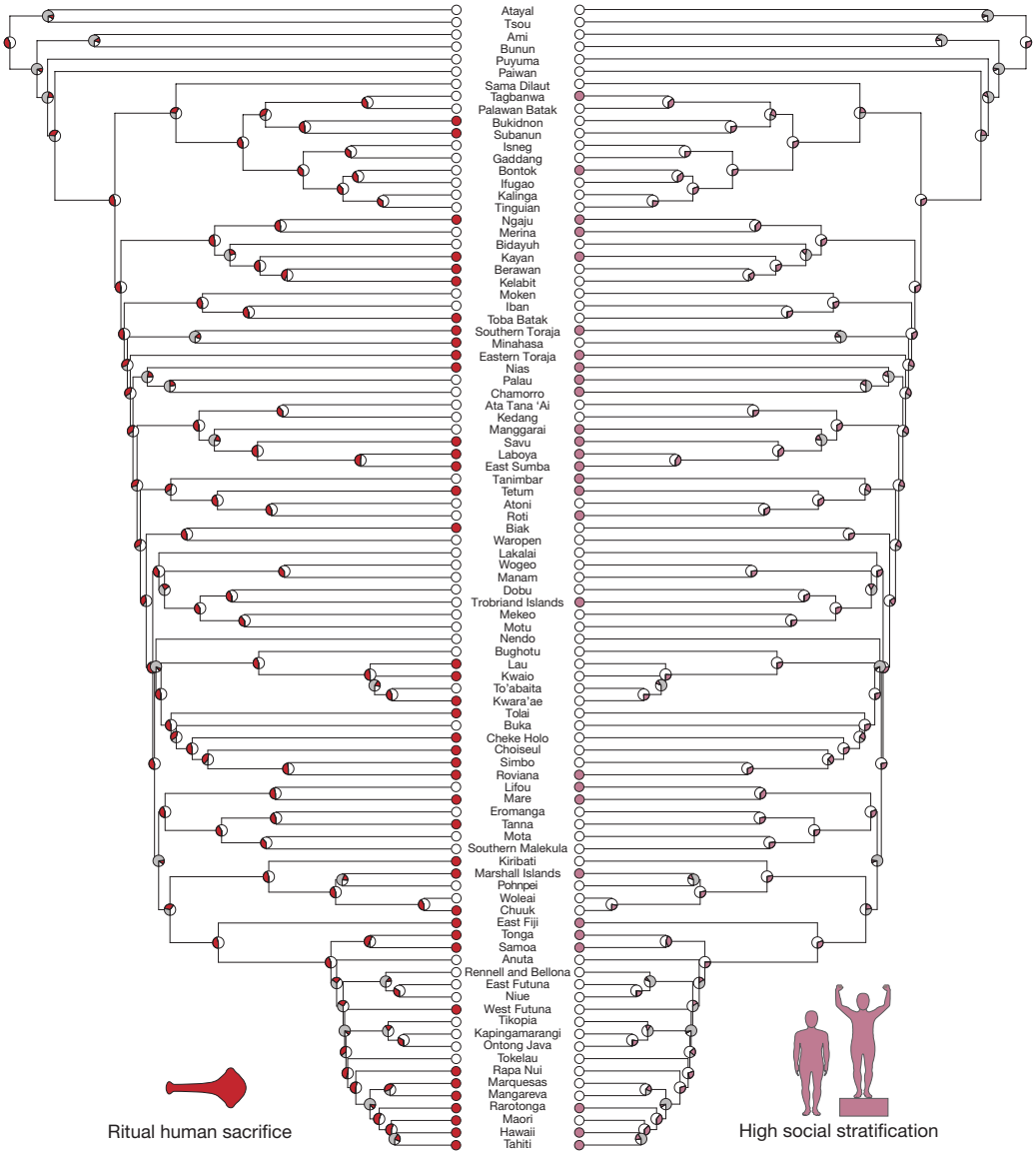


Figure 2 | Phylogenetic distribution of human sacrifice and high social stratification in Austronesia. Ancestral state reconstruction of human sacrifice and high social stratification on a maximum clade credibility consensus tree of 93 Austronesian languages. This analysis was run for 2×10^9 iterations and replicated three times. Pie charts at the nodes represent the probable ancestral state in the unconstrained dependent reversible-jump Markov chain Monte Carlo[23] analysis. Grey represents the proportion of our sample of 4,200 trees in which that node is absent.

slightly less likely than the models in the independent analysis, though not substantially so (BF = 1.34). The results from our second series of analyses indicates that human sacrifice increased the rate at which cultures with human sacrifice gain high social stratification, but did not function to stabilize high social stratification once it had arisen (in Fig. 1e, rate *d* is higher than rate *b*).

Taken together, our results provide strong evidence for the claim that human sacrifice played a powerful role in the construction and maintenance of stratified societies. Though human sacrifice was practiced in the majority of highly stratified societies in our sample, it was scarce in egalitarian societies, and we find that its effect depended on the level of stratification. Specifically, human sacrifice substantially increased the chances of high social stratification arising and prevented the loss of social stratification once it had arisen, yet was not found to increase social stratification in egalitarian societies. This is consistent with historical accounts that speculate that in order for human sacrifice to be exploited by social elites, there must first be social elites to exploit it[2,8]. In our ancestral reconstructions Proto-Austronesian culture is inferred to have had some level of social stratification (Extended Fig. 1), but not high social stratification (Fig. 2), and the most common changes inferred across our trees were the loss of social stratification in general, and the gain in high social stratification. We caution that the lack of support we find for human sacrifice sustaining high social stratification may be due to high social stratification having been rarely lost in the history of Austronesian cultures.

Experimental research indicates that while social inequality may foster group decision-making and efficiency[28], power hierarchies become unstable when they lack sanctioning status[29]. In Austronesian cultures human sacrifice was used to punish taboo violations[4], demoralise underclasses[27], mark class boundaries[3], and instil fear of social elites[27] — proving a wide range of potential mechanisms for maintaining and building social control. Throughout human history the practice of human sacrifice was often used by social elites as a display of power[2,8], intended to instil fear of the secular and supernatural consequences of transgressing ruling authority. While there are many factors that help build and sustain social stratification, human sacrifice may be a particularly effective means of maintaining and building social control because it minimizes the potential of retaliation by eliminating the

victim, and shifts the agent believed to be ultimately responsible to the realm of the supernatural[13].

Religion has long been proposed to play a functional role in society[19], and is commonly claimed to underpin morality. Recent evolutionary theories of religion have focused on the potential of pro-social and moral religious beliefs to increase cooperation[9,10]. Our findings suggest that religious rituals also played a darker role in the evolution of modern complex societies. In traditional Austronesian cultures there was substantial religious and political overlap, and ritualised human sacrifice may have been co-opted by elites as a divinely sanctioned means of social control[11,12,30]. The approach adopted in this paper demonstrates the way causal hypotheses about major transitions in human social organization can be tested by combining computational models and language phylogenies with a wealth of cultural and historical data. Unpalatable as it might be, our results suggest that ritual killing helped humans transition from the small egalitarian groups of our ancestors, to the large stratified societies we live in today.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 17 August 2015; accepted 25 January 2016.
Published online 4 April 2016.

1. Bremmer, J. N. *The Strange World of Human Sacrifice* (Peeters, 2007).
2. Carrasco, D. *City of Sacrifice* (Beacon Press, 1999).
3. Beatty, A. *Society and Exchange in Nias* (Claredon Press, 1992).
4. Burt, B. *Tradition and Christianity: the Colonial Transformation of a Solomon Islands Society* (Harwood Academic Publishers, 1994).
5. Del Castillo, B. D. *The History of the Conquest of New Spain.* (Univ. New Mexico Press, 2008).
6. Johnson, T. M. & Grimm B. J. *The World's Religions in Figures: an Introduction to International Religious Demography* (Wiley, 2013).
7. Winkelman, M. Political and demograpic-ecological determinants of institutionalised human sacrifice. *Anthropol. Forum* **24,** 47–70 (2014).
8. Turner, C. G. & Turner, J. A. *Man Corn: Cannibalism and Violence in the Prehistroic American Southwest* (Univ. Utah Press, 1999).
9. Norenzayan, A. *et al.* The cultural evolution of prosocial religions. *Behav. Brain Sci.* **39,** http://dx.doi.org/10.1017/S0140525X14001356 (2014).
10. Watts, J. *et al.* Broad supernatural punishment but not moralising high gods precede the evolution of political complexity in Austronesia. *Proc. R. Soc. B* **282,** http://dx.doi.org/10.1098/rspb.2014.2556 (2015).
11. Cronk, L. Evolutionary theories of morality and the manipulative use of signals. *Zygon* **29,** 81–101 (1994).
12. Marx, K. & Engels, F. *Karl Marx and Friedrich Engels: Collected Works* (International Publishers, 1975).
13. Girard, R., Hamerton-Kelly, R. G., Burkert, W. & Smith, J. Z. *Violent Origins: Walter Burkert, René Girard & Jonathan Z. Smith on Ritual Killing and Cultural Formation* (Stanford Univ. Press, 1987).
14. Burkert, W. *Creation of the Sacred: Tracks of Biology in Early Religions* (Harvard Univ. Press, 1998).
15. Price, B. J. Demystification, enriddlement, and Aztec cannibalism: a materialist rejoinder to Harner. *Am. Ethnol.* **5,** 98–115 (1978).
16. Harner, M. The ecological basis for Aztec sacrifice. *Am. Ethnol.* **4,** 117–135 (1977).
17. Gould, S. J. & Lewontin, R. C. The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proc. R. Soc. Lond. B* **205,** 581–598 (1979).
18. Wilson, E. O. *On Human Nature* (Harvard Univ. Press, 1978).
19. Coulanges, F. *The Ancient City: a study of Religion, Laws, and Institutions of Greece and Rome* (Lee and Shepard, 1877).
20. Flannery, K. & Marcus, J. *The Creation of Inequality: How Our Prehistroic Ancestors Set The Stage for Monarchy, Slavery, and Empire* (Harvard Univ. Press, 2012).
21. Wheatley, P. *The Pivot of the Four Quarters* (Edinburgh Univ. Press, 1971).
22. Dow, M. & Eff, E. Global, regional, and local network autocorrelation in the standard cross-cultural sample. *Cross-Cultural Res.* **42,** 148–171 (2008).
23. Pagel, M. & Meade, A. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *Am. Nat.* **167,** 808–825 (2006).
24. Watts, J. *et al.* Pulotu: database of Austronesian supernatural beliefs and practices. *PLoS ONE* **10,** e0136783 (2015).
25. Gray, R. D., Drummond, A. J. & Greenhill, S. J. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* **323,** 479–483 (2009).
26. Goodenough, W. Oceania and the problem of controls in the study of cultural and human evolution. *J. Polyn. Soc.* **66,** 146–155 (1957).
27. Kamakau, S. M. *Ka Po'E Kahiko: the People of Old* (Bernice P. Bishop Museum Press, 1968).
28. Ronay, R., Greenaway, K., Anicich, E. M. & Galinsky, A. D. The path to glory is paved with hierarchy: when hierarchical differentiation increases group effectiveness. *Psychol. Sci.* **23,** 669–677 (2012).
29. Lammers, J., Galinsky, A. D., Gordijin, E. H. & Sabine, O. Illegitimacy moderates the effect of power on approach. *Psychol. Sci.* **19,** 558–564 (2008).
30. Bellah, R. N. *Religion in Human Evolution: From the Paleolithic to the Axial Age* (Harvard Univ. Press, 2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We would like to thank K. Sterelny for feedback on an earlier version of the manuscript as well as M. Pagel and A. Meade for assistance with BayesTraits. We would also like to thank The John Templeton Foundation (28745), Templeton World Charity Foundation (0077), a Rutherford Discovery Fellowship (RDF-OUA1101), a PhD scholarship from the University of Auckland, and the Marsden Fund (UOA1104, VUW1321) for funding.

Author Contributions J.W. designed the study with Q.D.A., J.B. and R.D.G. J.W. and O.S. jointly created and coded the variables. J.W. performed the analyses with input from Q.D.A and R.D.G. J.W., O.S., Q.D.A., J.B. and R.D.G. reviewed the results and wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.W. (me@josephwatts.org).

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The investigators were not blinded during experiments and outcome assessment.

Data and coding. Human sacrifice was coded as present (present = 1) if there was evidence that members of the culture practiced the ritual killing of human beings, in a non-military context, for the sole or primary purpose of pleasing or appeasing a supernatural agent. Deaths that occurred during raids on enemy cultures, or non-ritual murders that resulted from interpersonal conflicts, were not considered to be human sacrifice. Human sacrifice was coded as absent (absent = 0) if ethnographic sources explicitly stated that human sacrifice was not practiced, or if there was no evidence of human sacrifice from a substantial description of the culture's religious practices. Building on an established classification system of social stratification in Polynesian cultures[31], we grouped Austronesian cultures into one of three categories. Cultures were coded as egalitarian (egalitarian = 1) if there was minimal or no potential for wealth and/or status to be inherited between generations. Cultures were coded as having moderate social stratification (moderate social stratification = 2) if pronounced intergenerational differences in wealth and/or status existed between social groups, but one or more of the following conditions was met: (a) social mobility was not restricted at any level, (b) differences in status and/or wealth were not associated with pronounced differences in living standards, and/or (c) the social groups in question were not clearly delineated. Finally, cultures were coded as highly stratified (high social stratification = 3) if pronounced intergenerational differences in wealth and/or status, associated with pronounced differences in living standards, existed between clearly delineated social groups and social mobility between two or more of the groups was restricted.

Cultures were the units of analysis in this study. We coded all the cultures from the Pulotu database[24] for which data on human sacrifice and social stratification were available, and that could be linked to a language on a phylogenetic tree[25]. Pulotu contains a diverse and broadly representative sample of Austronesian speaking cultures and information was collected on the traditional states of these cultures from periodicals, books and encyclopaedias (Supplementary Table 1). We coded cultures as they were before substantial influence by industrialized cultures and major world religions. This influence occurred through modernizing processes such as colonization, missionization, and trade. Sampling traditional Austronesian cultures has the advantages of reducing the effects of cultural diffusion and enabling us to test our hypotheses using a sample of cultures with a diverse range of religious and social structures[26]. The social structures of these cultures ranged from small kin-based groups to large, polities such as those found in Hawaii[27], meaning that that our sample is particularly well suited to testing hypotheses about the evolutionary transitions that occurred in early human civilizations[20]. In the collection of ethnographic data, priority was given to primary ethnographic materials, collected by ethnographers nearest to the time focus. Each culture was coded by two trained coders and the ethnographic sources used for each culture, as well as the coding decisions, can be found in the Supplementary Information. The first coder found and reviewed suitable ethnographic materials and coded each of the variables. The second coder reviewed this decision based on a review of the sources consulted by the first coder and a search for any additional materials. The coders' decisions were highly consistent. In rare cases of disagreement, a third coder was consulted and a clear agreement was reached between all coders.

Tree building. To model the ancestral history of cultures we used a sample of 4,200 trees from the posterior distribution of a Bayesian analysis of Austronesian basic vocabulary items (the detailed method is described in Gray, Drummond, and Greenhill[25]). The cultural history inferred by these language trees is corroborated by current genetic data[32], and what is known from the archaeological record about the sequence and timing of cultural expansions[25]. We pruned the original sample of 400 languages to 93, selecting those corresponding to cultures that were the subject of detailed ethnographic descriptions, while ensuring we sampled from all major cultural groupings and geographic regions.

Preliminary phylogenetic analyses. First, we used the multistate function in BayesTraits to test for patterns in the evolution of social stratification, without the influence of human sacrifice. While the discrete function in BayesTraits requires binary traits, the multistate function can be used to test how a trait with more than two states evolves[23]. In this analysis we tested three different models of evolution. The first was unconstrained so that any transition between states could occur. For example, cultures could transition directly from being egalitarian to either moderate or high social stratification. The second model of evolution was constrained so that stratification must be gained in steps from egalitarian to moderately stratified, to highly stratified, but could be lost in jumps. For example, a culture could transition directly from being highly stratified to being egalitarian, but not vice versa. The third model of evolution was constrained so that social stratification must be gained and lost in sequential steps. This means that for a culture to go from being egalitarian to highly stratified, or highly stratified to egalitarian, it must pass through a stage of being moderately stratified. We find that neither our second analyses that require cultures to gain social stratification in steps (BF = 0.08), or our third models that require the gain and loss of social stratification to occur in steps (BF = 1.28), were supported over the unrestricted models of evolution (Supplementary Tables 2–5). In the unconstrained analyses the mean transition rates between different states of social stratification were equal (Supplementary Table 2), and a range of different model were sampled in the posterior distribution (Supplementary Table 3). This suggests that cultures can transition freely between each different level of social stratification. These findings mean that in order to test how human sacrifice co-evolves with social stratification using BayesTraits, which requires binary traits, it is appropriate to group moderate and high stratification together as there could have been transitions directly to or egalitarianism from either form of stratification. These findings also mean that in order to test for the co-evolution of human sacrifice and high social stratification, it is appropriate to group egalitarianism with moderate social stratification as either could have transitioned directly to or from high social stratification.

We then tested for phylogenetic signal in our traits by using the phylo.d function in the R[33] package Caper[34] to calculate Fritz and Purvis' D[35], as well as whether this value differed significantly from what would be expected given no phylogenetic patterning, or under a Brownian model of evolution. We performed 1,000 permutations for each tree in our 4,200 tree sample, and we present the mean and standard deviation of these values across the sample of trees. A D statistic of 0 indicates that a trait is as phylogenetically conserved as would be expected under a Brownian model of trait evolution, while a value of 1 indicates that the distribution of the trait is not phylogenetically patterned[35]. Our results indicate that human sacrifice is highly conserved ($D = -0.03$, s.d. $= 0.09$), and that its distribution is not significantly different from what would be expected under a Brownian model of trait evolution ($P = 0.54$), but is significantly different from what would be expected if there were no phylogenetic signal ($P < 0.01$). Our results also indicate that social stratification ($D = 0.19$, s.d. $= 0.10$) and high social stratification ($D = 0.18$, s.d. $= 0.11$) are phylogenetically patterned. The distribution of both social stratification and high social stratification was significantly different from that expected if there were no phylogenetic signal ($P = 0.01$ and $P < 0.01$, respectively), and not significantly different from that which would be expected under a Brownian model of evolution ($P = 0.39$ and $P = 0.38$, respectively). The strength of phylogenetic signal means that the assumptions of standard non-phylogenetic methods are violated, and that phylogenetic methods are appropriate to account for the historical dependencies between cultures[36,37].

Co-evolution models. Co-evolutionary analyses were performed in the phylogenetic software package BayesTraits[23]. In order to inform our choice of priors for the MCMC analyses, we began by performing maximum likelihood (ML) analyses. Setting the number of optimisation attempts at 100 per tree, we calculated the mean transition rates for dependent and independent models across our sample of trees. For the models of human sacrifice and social stratification the mean transition rates ranged from 0.03 to 0.46. For the human sacrifice and high social stratification analyses the mean rates ranged from <0.01 to 0.39.

We then used the MCMC function in BayesTraits[23] to test for correlated evolution between traits. Using the MCMC function has the advantage of being able to test models of evolution across a sample of trees, rather than just one, which allows for phylogenetic uncertainty to be accounted for. We tested for co-evolution by comparing the likelihood of posterior distribution of dependent and independent analyses. Dependent analyses allow the evolution of one trait to depend on the state of another trait, and should be favoured when co-evolution has occurred. For example, the chance of a culture gaining high social stratification may be higher in cultures with human sacrifice than in cultures without. Independent analyses contain only models in which the evolution of one trait is independent of the other[23]. For example, the chance of a culture gaining high social stratification will be the same for cultures with and without human sacrifice. In order to avoid over-parameterizing the model, we used a reverse-jump method that minimises the number of rate parameters used by only adding additional rate parameters when they improved the fit of the model. We used a hyper-prior seeding from an exponential distribution, and used the results of the ML analyses to inform the range of this hyper-prior. For all of our co-evolutionary analyses we set the hyper-before range from 0 to 0.5. Each analysis was run for 2×10^9 iterations, with the first 10^9 removed as a burn-in period. At the end of each run we calculated the log marginal likelihood by running a stepping-stone sampler[38] across the posterior distribution of the analyses. This stepping-stone sampler used a beta (0.40, 1.00) distribution and was run for 100,000 iterations across 1,000 stones. In order to ensure consistency we ran each analysis three times and reported the mean values across the run. As can be seen in the Supplementary Information, for all analyses, each of the three runs converged on highly consistent values (Supplementary Tables 6–16). We calculated Bayes Factors as twice the difference between the

log marginal likelihood of the posterior distributions of each analysis. Following Raftery[39] we take Bayes factors of 0–2 as providing no support for the models in one posterior distribution over the models in another posterior distribution, Bayes factors of 2–5 as providing positive support for one posterior distribution over the other, a Bayes factor of 5 to 10 as strong support, and a Bayes factor over 10 as very strong support.

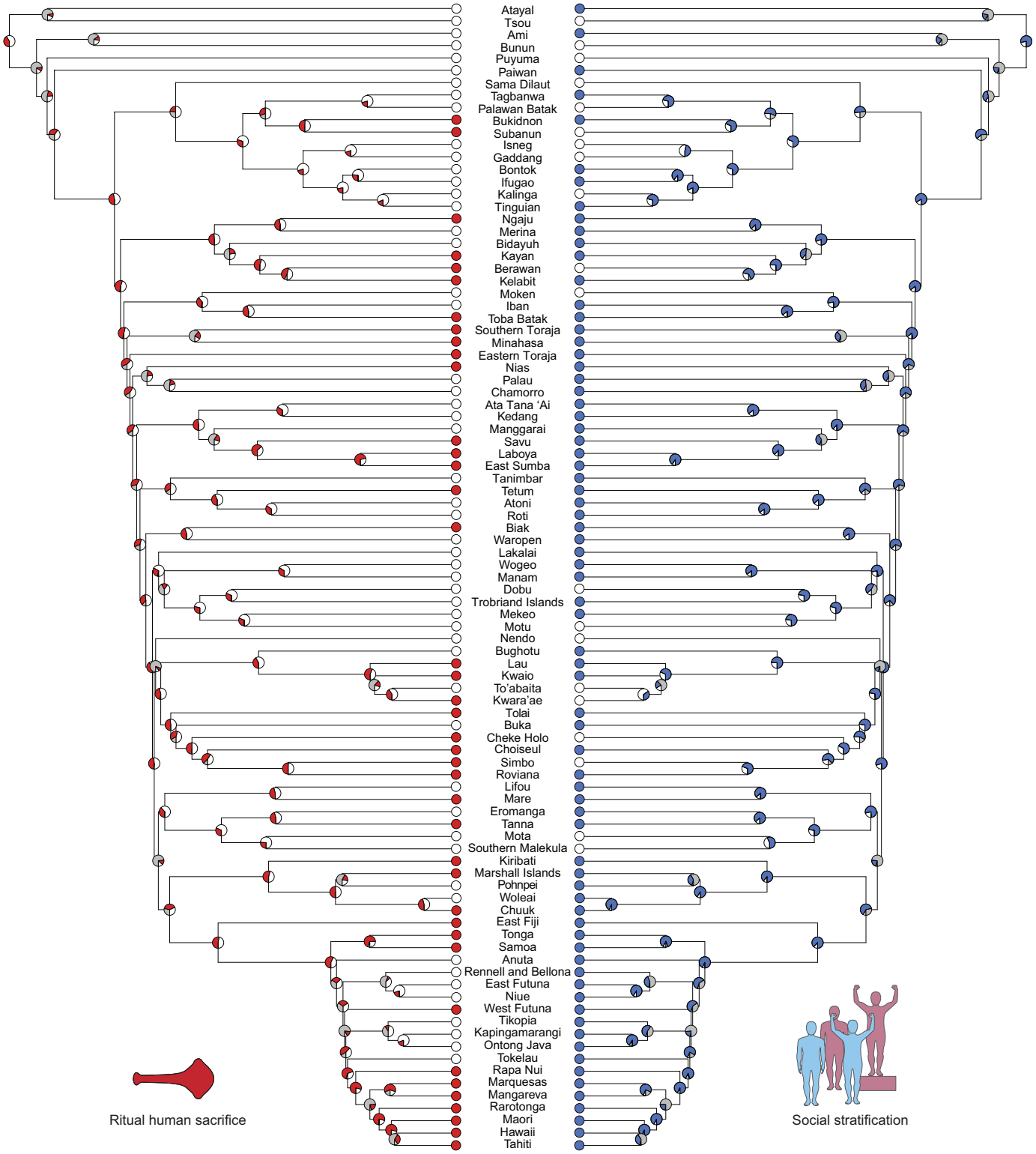
To test why the dependent model was favoured over the independent model, we performed follow-up analyses in which the dependent model was constrained. By constraining the dependent model, we could force the MCMC chain to sample a subset of the models in the unconstrained dependent analyses. For example, to test whether human sacrifice affects the rate at which cultures gain high social stratification we forced the dependent analyses to sample only models of co-evolution in which cultures with and without human sacrifice have an equal chance of gaining high social stratification (rates *b* and *d* in Fig. 1e can be set to be equal). If human sacrifice were to affect the rate at which cultures gain high social stratification, then we should expect the constrained models to fit substantially more poorly than the unconstrained dependent models.

Validation of co-evolutionary models. In order to test for sampling effects we validated our co-evolutionary models by randomly sub-sampling 50% of cultures from the full analyses, and then re-ran all co-evolutionary models with this sub-sample. This process was repeated with 10 different random sub-samples (summarized in Supplementary Tables 17–26). Despite the reduced power of the smaller sample, we find support for the same pattern of evolution in the majority of random sub-samples. For the co-evolutionary analyses of social stratification and human sacrifice we find substantial or strong support for the models in the dependent analysis over the models in the independent analysis, as well the same pattern of co-evolution as that found in our full analyses, in eight out of ten random sub-samples. One of the remaining two sub-samples found substantial support for the same pattern of co-evolution, but no single model structure accounted for over half of those sampled in the posterior distribution and the most commonly sampled model structure differed from that of the full analyses. In the other remaining sub-sample, the dependent model suggested the same pattern of co-evolution but was not substantially supported over the independent model. For the co-evolutionary analyses of high social stratification and human sacrifice, we find strong or substantial support for the models in the dependent analyses over the models in the independent analyses, as well as the same pattern of co-evolution as that found in our full analyses, in seven out of the ten randomly selected sub-samples. In the three remaining sub-samples, the dependent model indicated the same pattern of co-evolution as in our full analyses, but the dependent models were not substantially supported over the independent models. The results of these random sub-sampling analyses indicate that our findings are robust across a wide range of randomly selected sub-samples and that even after substantially reducing our power to detect correlated evolution we find support for the same relationship between social stratification and human sacrifice.

Recent simulation studies by Maddison and FitzJohn[40] have highlighted the potential for phylogenetic methods to lead to spurious correlations when one or more of the traits has undergone only a small number of evolutionary transitions such as one or two changes on a tree. As can be seen in Fig. 2 and Extended Fig. 1, both human sacrifice and social stratification are likely to have undergone a far greater number of changes than this, and these changes are distributed throughout Austronesia, indicating that the issues identified by Maddison and FitzJohn[40] do not apply to our study.

Figure construction. We created Fig. 2 and Extended Data Fig. 1 using a maximum clade credibility consensus tree from the full sample of trees using the software programme TreeAnnotator[41]. Trees were plotted using the plot.phylo function in the *R* package ape[42], and the node values were plotted using the nodelabels function. The probabilities assigned to the nodes of the tree were calculated by using the add-nodes function in BayesTraits[23], and represent the mean values assigned to nodes in the posterior distributions of the MCMC analyses. In these figures, grey was labelled 'phylogenetic uncertainty', and was used to illustrate the proportion of the trees in the sample for which that specific node of the consensus tree was absent.

31. Sahlins, M. D. *Social Stratification in Polynesia.* (Univ. Washington Press, 1958).
32. Ko, A. M. *et al.* Early Austronesians: into and out of Taiwan. *Am. J. Hum. Genet.* **94,** 426–436 (2014).
33. Team, R. Core. R: A Language and Environment for Statistical Computing. (R Found. Stat. Comput. 2015).
34. Orme, D. *et al.* Caper: Comparative Analyses of Phylogenetics and Evolution in R. R package version 0.5.2. http://cran.r-project.org/package=caper (2013).
35. Fritz, S. A. & Purvis, A. Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits. *Conserv. Biol.* **24,** 1042–1051 (2010).
36. Mace, R. & Pagel, M. The comparative method in anthropology. *Curr. Anthropol.* **35,** 549–564 (1994).
37. Mace, R. & Holden, C. J. A phylogenetic approach to cultural evolution. *Trends Ecol. Evol.* **20,** 116–121 (2005).
38. Xie, W., Lewis, P. O., Fan, Y., Kuo, L. & Chen, M.-H. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst. Biol.* **60,** 150–160 (2011).
39. Raftery, A. E. in *Markov Chain Monte Carlo in Practice* 163–188 (Chapman & Hall, 1996).
40. Maddison, W. P. & FitzJohn, R. G. The unsolved challange of phylogenetic correlation tests for categorical characters. *Syst. Biol.* **64,** 127–136 (2015).
41. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29,** 1969–1973 (2012).
42. Paradis, E., Claude, J. & Strimmer & K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20,** 289–290 (2004).



Extended Data Figure 1 | Phylogenetic distribution of human sacrifice and social stratification in Austronesia. Ancestral state reconstruction of human sacrifice and general social stratification on a maximum clade credibility consensus tree of 93 Austronesian languages. This analysis was run for 2×10^9 iterations and replicated three times. Pie charts at the nodes represent the probable ancestral state inferred in an unconstrained dependent reversible-jump Markov chain Monte Carlo[23] analysis. Grey represents the proportion of our sample of 4,200 trees in which that node is absent.

LETTER

doi:10.1038/nature17176

Post-invasion demography of prehistoric humans in South America

Amy Goldberg[1]*, Alexis M. Mychajliw[1]* & Elizabeth A. Hadly[1,2]

As the last habitable continent colonized by humans, the site of multiple domestication hotspots, and the location of the largest Pleistocene megafaunal extinction, South America is central to human prehistory[1–7]. Yet remarkably little is known about human population dynamics during colonization, subsequent expansions, and domestication[2–5]. Here we reconstruct the spatiotemporal patterns of human population growth in South America using a newly aggregated database of 1,147 archaeological sites and 5,464 calibrated radiocarbon dates spanning fourteen thousand to two thousand years ago (ka). We demonstrate that, rather than a steady exponential expansion, the demographic history of South Americans is characterized by two distinct phases. First, humans spread rapidly throughout the continent, but remained at low population sizes for 8,000 years, including a 4,000-year period of 'boom-and-bust' oscillations with no net growth. Supplementation of hunting with domesticated crops and animals[4,8] had a minimal impact on population carrying capacity. Only with widespread sedentism, beginning ~5 ka[4,8], did a second demographic phase begin, with evidence for exponential population growth in cultural hotspots, characteristic of the Neolithic transition worldwide[9]. The unique extent of humanity's ability to modify its environment to markedly increase carrying capacity in South America is therefore an unexpectedly recent phenomenon.

Genetic, archaeological and linguistic evidence suggest the first Americans descended from ancestral Siberians. A small population or populations crossed the land bridge connecting Asia and Alaska between 15 ka and 30 ka, reaching southern South America by at least 14.5 ka [1–6]. Yet the peopling of South America and the relative effects of climate and culture on early and mid-Holocene population dynamics remains unclear[2–5,8,10,11], particularly during the rise of agriculture worldwide.

More generally, with its recent and rapid colonization, South America provides a unique opportunity to study the colonization behaviour and population growth dynamics of modern humans. South America is exceptional as it was peopled during a single wave of invasion over a narrow time window, with limited later migration until historic times[5,6]. Therefore, it provides a continental view of human population history unavailable elsewhere.

To discern the dynamics of the South American human populations through the Holocene, we compiled a database of 5,464 radiocarbon (^{14}C) dates from 1,147 archaeological sites associated with human occupation (Fig. 1, Supplementary Data 1). We confined our analyses to dates ranging from late-Pleistocene colonization to 2 ^{14}C ka. Dates are calibrated unless specified as radiocarbon dates, ^{14}C ka. We use two proxies for human population size: the probability density of summed calibrated radiocarbon dates (SCPDs)[12–15], and the number of occupied archaeological sites over time[13–16]. After applying quality control filters and pruning the data to prevent oversampling, we are left with a set of 2,576 approximately independent merged occupation events for SCPD analyses.

Archaeological evidence of the initial peopling of South America is scarce, with many of the earliest sites found in Patagonia, despite geographic constraints and genetic data suggesting a north-to-south colonization route[1–6]. Indeed, once there is evidence of occupation, humans were already dispersed throughout the continent. Although humans are geographically widespread, initial site and date densities suggest low overall population sizes (Fig. 2a and Extended Data Figs 1 and 2). However, because sea levels were ~100 m lower during the interval of colonization and a coastal route of migration has been hypothesized, the earliest archaeological sites may be on the continental shelf that is now underwater[3–6].

Small human population sizes are almost immediately followed by a rapid increase in the density of radiocarbon dates and the number of occupied sites from 13 ka to 9 ka (Figs 2 and 3). At ~9 ka, the SCPD stabilizes, oscillating around a constant mean for ~4,000 years (Fig. 3). During this period, we find a peak in the SCPD occurring at ~11 ka, as well as recurring peaks from 9 ka to 5.5 ka.

It is tempting to interpret the initial peak around 11 ka and associated abrupt decline as evidence of overshooting carrying capacity after rapid colonization and correction to a carrying capacity over time. Such an overshoot is typical of species growing in a new environment, and is often linked to the rapid demise of a main food resource, perhaps indicative of the South American megafaunal extinction in this case[7,17]. Notably, our simulations suggest the recurring mid-Holocene peaks and troughs (9 ka to 5.5 ka) cannot be explained by calibration artefacts alone, but rather represent a dynamic equilibrium with oscillation around a carrying capacity[18] (Extended Data Figs 3 and 4). After 4,000

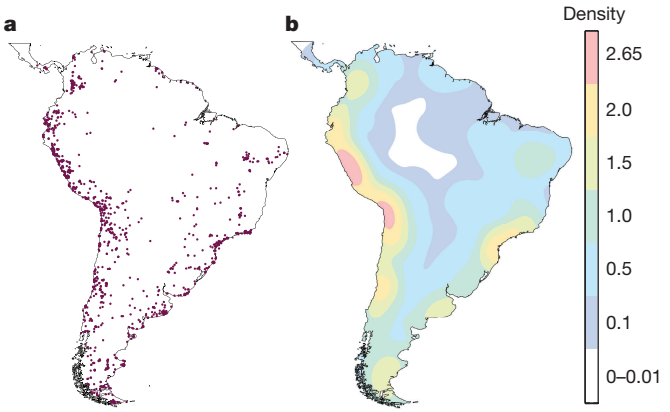


Figure 1 | Archaeological sites in database. a, Location of sites with radiocarbon dates with both laboratory numbers and latitude and longitude coordinates (N=1,147). **b**, Kernel-smoothed sampling intensity for the number of sampled sites used in analyses, with a search radius of 660 km. Density is measured by square decimal degree. Plotted in ArcGIS using the South America Albers Equal Area Conic Projection (standard parallels: −5, −42 degrees).

[1]Biology Department, Stanford University, Stanford, California 94305, USA. [2]Woods Institute, Stanford University, Stanford, California 94305, USA.
*These authors contributed equally to this work.

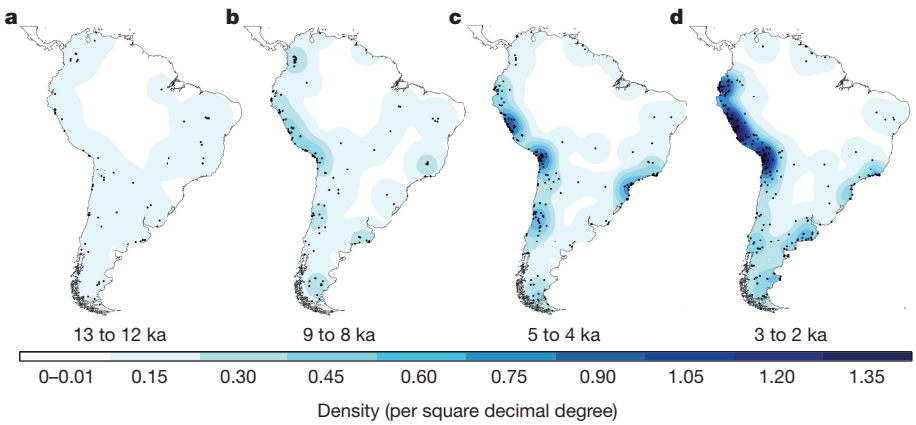


Figure 2 | Kernel density maps of occupied area. a–d, For 13 ka to 12 ka (**a**), 9 ka to 8 ka (**b**), 5 ka to 4 ka (**c**), and 3 ka to 2 ka (**d**), the kernel density smoothing of occupied area based on sites occupied in 1,000-year bins, with site locations, plotted in ArcGIS.

years of a dynamic equilibrium, there was renewed population growth, which was not followed by levelling off in our data set.

Kernel density maps of the land occupied by humans from ~14 ka to 2 ka suggest clear patterns in known centres of growth, with the earliest and highest population growth patterns seen on the Pacific coastline in Peru, Chile and Ecuador. To a lesser extent, we also witnessed centres of growth in Patagonia and the Brazilian Coastal Strip during the later mid-Holocene (Fig. 2 and Extended Data Figs 1 and 2). Lack of signal in certain regions, such as the Amazon, may reflect poor preservation of material or a lack of archaeological sampling, rather than an absence of people (Fig. 1b). South American population size dynamics show substantial spatial and temporal heterogeneity, which must be accounted for in models of human biological and cultural evolution.

For a first approximation of the overall growth in population size, we compared the mean density of the SCPD from 13.5 ka to 2.5 ka. We estimate that population size increased by roughly 1,000-fold during this time. To quantitatively analyse population growth rates and trends, we considered three general models of long-term human population growth: (1) exponential, (2) logistic, and (3) climate-mediated (Extended Data Table 1). Human populations are frequently modelled using exponential growth, uninhibited growth at a constant rate over time[12,19,20]. Recent discussion has focused around the explosive growth of the last few hundred years of global human populations[20–22], but the applicability of an exponential growth model for the majority of human history remains unclear. An alternative to unconstrained growth is density-dependent or logistic growth. As populations expand into new and favourable habitats, they often experienced rapid growth followed by a gradual decline in growth rate as they used resources and approached carrying capacity. This logistic growth model has been demonstrated theoretically and empirically in a variety of mammalian, avian, plant and bacterial populations[22–25]. A final model for population dynamics—climate-mediated growth—is one where the resources are determined by top-down controls characteristic of fluctuating environments[10,26].

We found trajectories of exponential growth and/or climate-mediated growth to be poor fits for the long-term human population size in South America (Fig. 4 and Extended Data Table 1). Rather, under a likelihood-style framework, the best fit was a two-phase model with density-dependent growth from initial peopling through 5.5 ka, followed by a recent phase of exponential growth from 5.5 ka to 2 ka (Fig. 4; minimum ΔBIC = 11.7, Schwarz weight >0.99). This model of growth is consistent with our kernel density analyses, and is distinct from the curvilinear shapes seen in SCPDs from North America, Europe, and Australia[12,14,15].

Under the model, initial growth was rapid ($r = 0.131\%$ per year, 3.33 to 3.74% per generation), with the population doubling every 19 to 21 generations (for a generation time between 25 to 28 years), followed by a relatively constant population size from 9 ka to 5.5 ka. Early growth

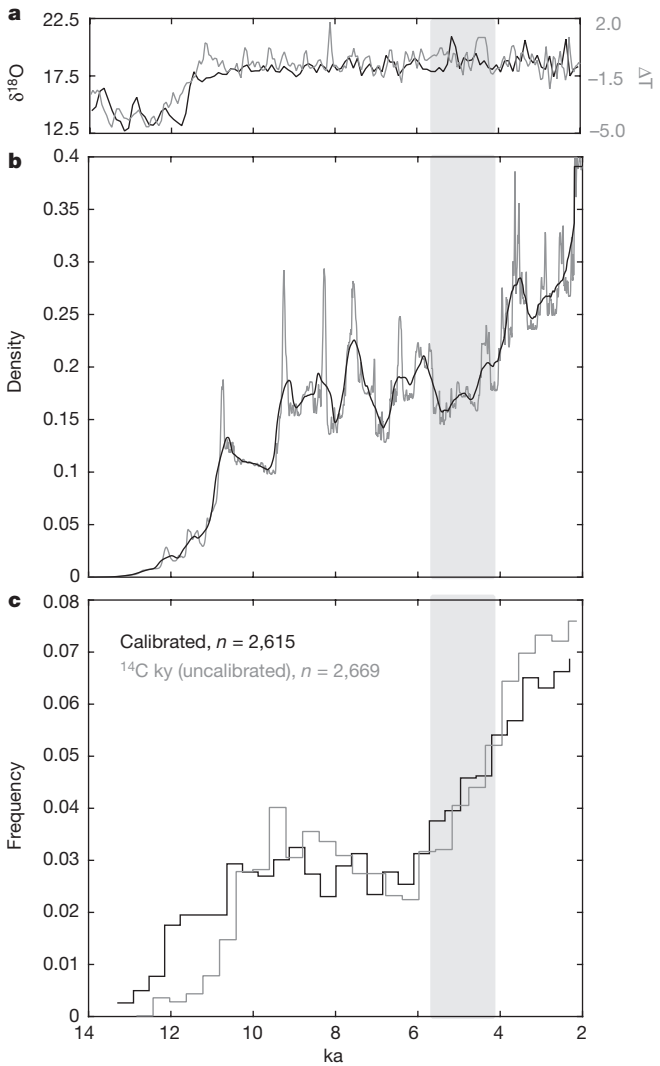


Figure 3 | Climatic and radiocarbon records for 14 ka to 2 ka. a, The 100-year averages of oxygen isotope data, $\delta^{18}O$ (‰), from Sajama, Bolivia ice core (black) and the difference in temperature from recent time average reconstructed from Vostok, Antarctica ice core (grey). **b**, SCPD for South American continent (grey), with a 400-year moving average (black). **c**, Frequency of occupied sites over time in 400-year time bins using the median of calibrated dates (95.4% distribution, black) and the frequency of occupied sites over time in 400-year time bins using uncalibrated radiocarbon dates (grey). For all panels, the grey bar indicates estimated time of demographic phase shift to renewed growth.

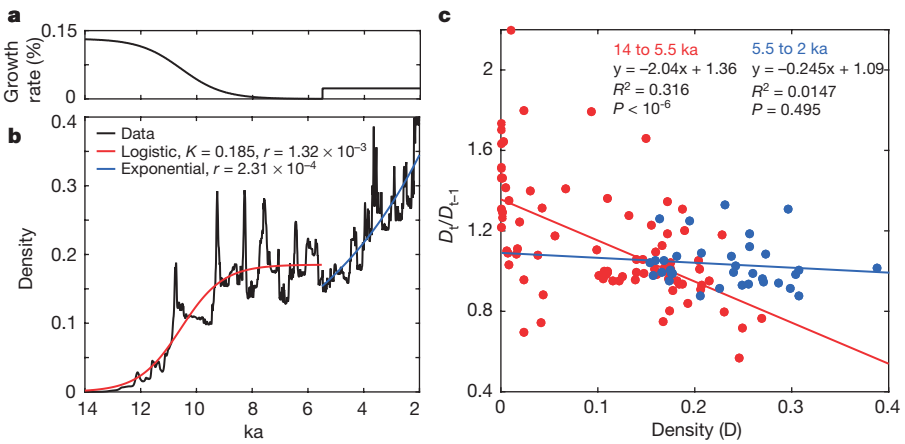


Figure 4 | A two-phase model for the human invasion of South America. **a**, Yearly growth rate (percentage) of the population over time under the model. **b**, SCPD for South American continent, with the best-fit model of population size over time: logistic growth (red) from colonization to 5.5 ka, and exponential growth from 5.5 ka to 2 ka. **c**, A linear model for density-dependent growth from 14 ka to 5.5 ka (red), with the rate of change of density significantly negatively correlated with current density, and from 5.5 ka to 2 ka (blue), where there is no evidence for density-dependent growth. Rather, for 5.5 ka to 2 ka, the constant rate is characteristic of exponential growth. The density and rate of change of density in the SCPD is calculated in 100-year bins, with each point representing the mean density of that point ± 50 years.

rates for South America are approximately an order of magnitude higher than archaeological and genetic estimates from the same period in Eurasia[18,20,27], and are on par with or larger than genetic estimates of explosive growth during the past 140 generations[21].

Before the mid-Holocene, human colonization of South America resembled the population size dynamics of a typical invasive species[24,25], with rapid initial expansion and resource-limited growth over time. However, after thousands of years with a dynamic equilibrium, South American populations experienced renewed growth, indicating an increase in carrying capacity. This increase is a sharp departure from animal models of population growth because it does not correlate with an amelioration of climate or known environmental change favourable to humans (Fig. 3). The growth rate in the second phase is estimated as $r = 0.579$ to 0.649% per generation.

We compared expansion dynamics of human populations in South America to a general density-dependence model for invasive species by comparing the density of radiocarbon dates to the rate of change of radiocarbon dates adapted from[24,25]. Fitting a linear model, the rate of population growth was inversely proportional to the current population size. We found a significant negative relationship for the period of inferred logistic growth. Conversely, the relationship was not significant during the period from 5.5 ka to 2 ka, consistent with exponential growth (Fig. 4c). The differing behaviour of the two linear models is additional evidence for two distinct phases of population dynamics.

Under our model, estimates for the yearly growth rate over time (Fig. 4a) can be used to infer population size (Extended Data Fig. 5). For an initial size of one thousand, this estimate of a ~600-fold increase is lower than, but consistent with, that from the ratio of radiocarbon dates at the start and end time periods alone. Together they give a range of approximately 615,000 to 1,000,000 people in South America by 2 ka, with more than half of the total population growth occurring between 5.5 ka to 2 ka.

Renewed growth is evidence of a change in carrying capacity in South America, with two, possibly interacting, causes: (1) climatic change, or (2) cultural or technological change. Palaeoclimate proxies do not support the hypothesis that the renewed growth was triggered by climatic change. Although the initial warming period (~12–11 ka) corresponds to an increase in population size during colonization, oxygen isotope levels fluctuate around constant mean through most of the Holocene (Fig. 3a). Palaeoclimatic and vegetation studies suggest climatic change during ~6 ka to 4 ka is region-specific and poorly understood, with no known global mechanisms for change[4,8,28]. Many of the most culturally developed and largest populations of the mid-Holocene overlap with harsh, dry climates, such as the coastal Chinchorro culture, and the northern Las Vegas culture[4,8] (Fig. 2 and Extended Data Fig. 6). Conversely, in Patagonia the population growth is more modest, consistent with lower levels of sedentism and domestication through the later Holocene[4,8].

South American population density during the mid-Holocene is in the lower range of estimates for the density of worldwide hunter-gatherer populations[27]. As populations grew in the mid-Holocene, formally isolated populations may have interacted more and, through trade and the spread of innovations, acted as a buffer towards the environment increasing the carrying capacity. This pattern can be cyclical, with larger populations leading to more opportunity for innovation and, in turn, the capacity for even larger populations, as well as unique interactions between culture and local environments[11,29]. We estimate population density increased threefold from 5.5 ka to 2 ka. The continental estimates for population density are low for agricultural populations[27], perhaps owing to regional variation in levels of agriculture[4,8].

South America had no known single cohesive culture comparable to Clovis or extensive evidence for long-range material and cultural sharing, likely due to divergent environments, geographic barriers to gene flow, and low population density[4,8]. As such, domestication was slow, involving long periods with a handful of domesticated plants and few animals supplementing hunting and gathering, starting in the northwest approximately 9 ka, with multiple domestication hotspots across the continent. This pattern is a sharp contrast to the European Neolithic, characterized by a somewhat constant rate of agricultural expansion from a singular Middle Eastern origin[9,30]. Even with vastly different domestication patterns and histories, South America surprisingly shows the same demographic transition characteristic of the European Neolithic[9].

Despite the long history of supplementary domestication, including staple crops such as squash, peppers and maize[4,8], the renewed growth did not begin until the shift to a predominantly sedentary and agricultural subsistence. This process occurred throughout the continent between 5.5 ka and 3.5 ka, though it did not fully reach Patagonia, as this region likely then, as now, was not amenable to agriculture. The timing of this transition overlaps well with our estimated time of renewed population expansion (5.7 ka to 4.1 ka). Stable food sources, the need for human labourers, coupled with increased parental availability lead to shorter birth intervals and larger populations in agricultural societies[9].

Humans are unique in the extent of their ability to manipulate their environments, and thus change local carrying capacity. Yet, we demonstrate that human prehistoric growth behaviour in fact resembles that of most animal populations—until the rise of agriculture. That is, although humans have been manipulating their environments through tool use and supplementary domesticated crops for millennia, it was not until the rise of widespread sedentism and agriculture that humans gained our distinctive ability to increase local and global carrying capacity.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 3 November 2015; accepted 26 January 2016.
Published online 6 April 2016.

1. Meltzer, D. J. *First Peoples in a New World: Colonizing Ice Age America* (Univ. California Press, 2009).
2. Raghavan, M. *et al.* Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* **349,** aab3884 (2015).
3. Dillehay, T. D. Probing deeper into first American studies. *Proc. Natl Acad. Sci. USA* **106,** 971–978 (2009).
4. Moore, J. D. *A Prehistory of South America: Ancient Cultural Diversity on the Least Known Continent* (Univ. Press of Colorado, 2014).
5. Rothhammer, F. & Dillehay, T. D. The late Pleistocene colonization of South America: an interdisciplinary perspective. *Ann. Hum. Genet.* **73,** 540–549 (2009).
6. Reich, D. *et al.* Reconstructing Native American population history. *Nature* **488,** 370–374 (2012).
7. Barnosky, A. D., Koch, P. L., Feranec, R. S., Wing, S. L. & Shabel, A. B. Assessing the causes of Late Pleistocene extinctions on the continents. *Science* **306,** 70–75 (2004).
8. Silverman, H. & Isbell, W. *Handbook of South American Archaeology* (Springer, 2008).
9. Bocquet-Appel, J. P. & Bar-Yosef, O. (eds) *The Neolithic Demographic Transition and its Consequences* (Springer, 2008).
10. Sandweiss, D. H., Solís, R. S., Moseley, M. E., Keefer, D. K. & Ortloff, C. R. Environmental change and economic development in coastal Peru between 5,800 and 3,600 years ago. *Proc. Natl Acad. Sci. USA* **106,** 1359–1363 (2009).
11. Marquet, P. A., Santoro, C. M., Latorre, C., Standen, V. G. & Abades, S. R. Emergence of social complexity among coastal hunter-gatherers in the Atacama Desert of northern Chile. *Proc. Natl Acad. Sci. USA* **109,** 14754–14760 (2012).
12. Shennan, S. *et al.* Regional population collapse followed initial agriculture booms in mid-Holocene Europe. *Nature Commun.* **4,** 2486 (2013).
13. Rick, J. W. Dates as data: an examination of the Peruvian preceramic radiocarbon record. *Am. Antiq.* **52,** 55–73 (1987).
14. Williams, A. N. The use of summed radiocarbon probability distributions in archaeology: a review of methods. *J. Archaeol. Sci.* **39,** 578–589 (2012).
15. Peros, M. C., Munoz, S. E., Gajewski, K. & Viau, A. E. Prehistoric demography of North America inferred from radiocarbon data. *J. Archaeol. Sci.* **37,** 656–664 (2010).
16. Mellars, P. & French, J. S. Tenfold population increase in Western Europe at the Neandertal-to-modern human transition. *Science* **333,** 623–627 (2011).
17. Barnosky, A. D. *et al.* The variable impact of Late-Quaternary megafaunal extinction in causing ecological state shifts in North and South America. *Proc. Natl Acad. Sci. USA* **113,** 856–861 (2016).
18. Boone, J. L. Subsistence strategies and early human population history: An evolutionary ecological perspective. *World Archaeol.* **34,** 6–25 (2002).
19. Haub, C. How many people have ever lived on earth? *Popul. Today* **23,** 4–5 (1995).
20. Keinan, A. & Clark, A. G. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336,** 740–743 (2012).
21. Gazave, E. *et al.* Neutral genomic regions refine models of recent rapid human population growth. *Proc. Natl Acad. Sci. USA* **111,** 757–762 (2014).
22. Cohen, J. E. *How Many People Can The Earth Support?* (W.W. Norton & Company, 1995).
23. Kingsland, S. The refractory model: the logistic curve and the history of population ecology. *Q. Rev. Biol.* **57,** 29–52 (1982).
24. Sibly, R. M. & Hone, J. Population growth rate and its determinants: an overview. *Phil. Trans. R. Soc. B* **357,** 1153–1170 (2002).
25. Arim, M., Abades, S. R., Neill, P. E., Lima, M. & Marquet, P. A. Spread dynamics of invasive species. *Proc. Natl Acad. Sci. USA* **103,** 374–378 (2006).
26. Kelly, R. L., Surovell, T. A., Shuman, B. N. & Smith, G. M. A continuous climatic impact on Holocene human population in the Rocky Mountains. *Proc. Natl Acad. Sci. USA* **110,** 443–447 (2013).
27. Hassan, F. A. *Demographic Archaeology in Studies in Archeology* (Academic Press, 1981).
28. Anderson, D. G., Maasch, K. & Sandweiss, D. H. (eds) *Climate Change and Cultural Dynamics: a Global Perspective on Mid-Holocene Transitions* (Academic Press, 2011).
29. Henrich, J. Demography and cultural evolution: how adaptive cultural processes can produce maladaptive losses: the Tasmanian case. *Am. Antiq.* **69,** 197–214 (2004).
30. Pinhasi, R., Fort, J. & Ammerman, A. J. Tracing the origin and spread of agriculture in Europe. *PLoS Biol.* **3,** e410 (2005).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank A. Barnosky, J. Kang, N. Rosenberg, E. Lindsey, N. Villavicencio, L. Frishkoff, K. Solari, J. Hsu, and H. Frank for conversations and feedback on the paper, as well as E. Jewett and M. Edge for discussions of statistical methods. We also greatly appreciate assistance from the Stanford Geospatial Center. This work was stimulated by discussions of an NSF-funded group (EAR 1148181) comparing the timing of megafaunal extinctions with archeological and palaeoenvironmental data in South America. We acknowledge support from NSF grant BCS 1515127, as well as NSF Graduate Research and ARCS fellowships to A.G., a Stanford Interdisciplinary Graduate Fellowship to A.M.M, and a Gabilan Fellowship to E.A.H. Radiocarbon data and associated information are available in the Supplementary Information.

Author Contributions A.G. conducted analyses and wrote the first draft of the paper. A.M.M. collected and collated the database, and conducted ArcGIS analyses. E.A.H. advised the analyses and initiated the project. All authors interpreted results, and contributed to framing and editing of the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.G. (agoldb@stanford.edu).

METHODS

Data reporting. No statistical methods were used to predetermine sample size.

Data, calibration, and quality control. Our data set of radiocarbon dates from archaeological sites in South America was gathered from previously published journal articles and online resources, including English, Spanish, French, and Portuguese publications (Supplementary Data 1). We included dates at least 2,000 ^{14}C years old. We follow the convention of refs 12 and 30, only excluding dates that have been rejected by the field, such as outliers for early colonization (before 12,780 ^{14}C years ago). This is expected to generate increased noise in the sample, but is likely outweighed by the increased sample size as the noise is approximately randomly distributed in the data set. We only included dates that have been associated with human activity, culture, or remains. No single sources of continent-wide databases for radiocarbon dates over this time period exist; therefore there are likely to be some omissions. Additionally, sea levels have risen since the first peopling of South America, therefore early coastal sites are not sampled.

Locations within 0.1 degrees latitude or longitude from each other are considered a single site. Sites were geo-referenced where appropriate. We excluded dates without an identifiable lab number or latitude and longitude for analyses. To buffer the effects of oversampling without substantial information loss, we considered a pruned version of the data set. Using the OxCal[31] R_combine command, we merged radiocarbon dates that occur within 200-year intervals at site into a single normalized probability density of calibrated calendar dates[12,15]. The SCPD for the pruned data are highly correlated ($R^2 = 0.971$, Pearson) with that of the unpruned data. Indeed, the unpruned data set provides stronger support for our two-phase model, with higher late Holocene growth rates. For all subsequent analyses and those in the main text, we use the pruned data set and assume independence of merged dates.

After these quality controls, 2,576 independent merged 'dates' at 1,147 unique sites. Radiocarbon dates and their associated errors were calibrated using OxCal 4.2 (ref. 31) and the most recent Southern Hemisphere calibration curve, SHCal13 (32).

Inclusion of shell dates. Our database includes 597 radiocarbon dates from marine shells or otoliths. Calcium carbonates from marine shells and otoliths require special calibration due to the marine reservoir effect, which results from temporal lags in carbon mixing within ocean waters. Reservoir effects can vary on local scales due to upwelling and circulation patterns, necessitating correction with both local and global corrections. We used published radiocarbon dates that were corrected for the local reservoir with a Delta*R* value. These dates were then calibrated with the Marine13 Curve[33] as is standard practice. Large coastal cultures throughout the continent are known to use marine resources heavily[4,8]; exclusion of shell dates would bias our results. Other studies of South American demography also opt to include marine shells in their analyses[34].

Palaeoclimate data. We used paleoclimatic data from two different sources to show general trends over large spatial scales. We consider data from ice cores sampled in Sajama, Bolivia[35], and Vostok, Antarctica[36].

Kernel density estimation. To visualize changes in site density and location through time, we created bivariate kernel density maps using the ArcGIS 10.2.2 (ref. 37) spatial analyst toolbox. We considered sites that are occupied in 1,000-year time bins, including only a single date per site within each period. Similar analyses using 500-year time bins did not change spatiotemporal patterns, and the larger archaeological sample sizes of 1,000-year time bins provide a clearer image of change through time. Thus we present the results of 1,000-year time bins here.

Kernel density estimation is widely used in spatial analyses, in which data in the form of points is used to create a density surface[38], based off smoothing of the surface near observed data and convolution of overlapping surfaces. We used an Epanecknikov kernel process[39], which more effectively minimizes mean integrated square error as compared with the alternative Gaussian kernel process[38,40]. Magnitudes are smoothed using a bilinear interpolation.

To prevent over or under smoothing, we consider two methods to choose the search radius (bandwidth), a scaling factor that extends from each data point and determines the width of the probability mass as it spreads from that point[40]: (1) independent best-fit search radius for each time slice, and (2) mean of independent best-fit radiuses applied to all time-slices.

For a given time period, with sample size n, the best-fit search radius, R, is given by[37]

$$R = 0.9n^{-1/5}\ \min(D_s,\ \ln(2)^{-1/2}D_m),$$

where D_m is the median of distances from the mean centre of all points in the data set, and D_s is the standard distance between those points.

We first consider the best-fit radius for each time period, which reflects differences in sampling and uncertainty. The radiuses ranged from 563 km to 1145 km, with the all but the oldest two time bins falling within a 560–700 km range, consistent with estimates for North America[41]. With these best-fit radiuses, we found a density range of 0.03 to 1.29 magnitude per square decimal degree.

We compared results from time-specific radiuses to a single radius for all time slices. We use the mean of all best-fit radiuses, 660 km, without outliers as a single search radius for all time periods[41] (Extended Data Fig. 2). Other values for the radius considered (500, 600, 800,1000) showed the same qualitative patterns. We used a density interval of 0.15, with nine discrete contours from 0.01–0.15 to 1.20–1.35 measured as magnitude per square decimal degree.

Summed probability distributions of calibrated dates and biases. *SCPDs as a proxy for population size*. The density of radiocarbon dates over time is an extensively used tool to estimate relative levels of population size[12–15,42]. We calculated the summed probability distributions of the calibrated radiocarbon dates (SCPDs), or the sum of the distributions for each calibrated date, using the *sum* function in OxCal at the two-sigma level. That is, an SCPD calibrates each merged radiocarbon date independently, which produces a set of 2,576 probability distributions of calendar dates. These calendar dates were then treated as independent random variables and their distributions summed to give a single convolution of all dates, the SCPD.

SCPDs are beneficial because they include the uncertainty associated with calibrated dates by considering each date as a distribution rather than a point estimate. The convolution of the distributions of each calibrated date into a single SCPD requires the assumption that the dates are independent. The pruning of dates described above into 200-year bins decreases autocorrelation.

We made two additional assumptions: (1) that the amount of material dated is approximately proportional to the population size, and therefore the density of material is a proxy for population size, and (2) that the amount of material dated is approximately proportional to the amount of material initially present, that is, that sampling and taphonomic biases are not large and systematic. These assumptions can be considered reasonable because independent lines of evidence demonstrate strong correlations between SCPDs and multiple other proxies of population size over time, suggesting that SCPDs reflect demographic events. Specifically, SCPDs closely track the number of archaeological sites, density of artefacts, palaeodemographic and burial histories, and environmental-use records[14,43–45], though one must interpret these with care[12–15,42,46]. While sampling biases in excavation and dating, as well as sampling of the literature to make the database, can influence the shape of the SCPD, our data set is large enough that we assume it approximates a random sampling, following recommendations in the literatue[14,42,47,48].

SCPDs assume a constant relationship between population size and the density of radiocarbon dates. Changes in SCPD may be the result of changes in the way populations distribute or leave material. In these cases, the SCPD may not reflect population size directly, but is still providing important information about changing subsistence strategy or culture. During the time period considered, most of the continent shifted from nomadic hunter-gatherers to sedentary populations. Owing to both field sampling and our quality control filters, which bias towards uniquely occupied sites[15,44], hunter-gatherers are likely to leave proportionally more radiocarbon dates. For example, pre-Neolithic hunter-gatherers lived approximately 25–30 individuals to a site, with more sedentary populations in the region often reaching 200, and occasionally reaching 400 individuals[9,27].

Taphonomic bias. Taphonomic degradation of material as a function of time is suggested to be systematic, introducing bias. Surovell *et al.*[49] provide a correction to the SCPD for taphonomy as a function of date based on volcanic records[14,15]. However, a singular correction is unfit for diverse climates on the scale of the South American continent[8,15,28]. The extensive preservation of the Atacama Desert cannot be considered in the same correction as dates from the highly degrading Amazon rainforest.

Additionally, we argue that owing to the particular emphasis on sampling earlier dates in the region because it coincides with first peopling and our quality controls such as the 200-year binning, taphonomic bias is a mild problem for our data[15]. Taphonomic bias should also effect dates more than sites, yet we see the same trend in the density of radiocarbon dates and two measures of site occupation.

The SCPD for South America presented here does not follow the curvilinear shape typical of strong taphonomic biases, where the frequency of dates is approximately exponentially related to the date, suggesting a weaker effect of taphonomy on curve shape. For example, during the 4,000-year period of approximately constant frequency of dates (9 to 5 ka), it is unreasonable to assume that the population size was decreasing at the approximate rate of taphonomic degradation over such a long period in diverse environments across the continent.

We acknowledge that some bias will exist in our data set, particularly due to differential sampling or excavation efforts, and geographic variation in preservation. For example, modern climate and vegetation plays a role in the lack of dates from the Amazon region (both sampling efforts and degradation), but may not be indicative of the absence of people. Therefore, it is not possible to draw conclusions from the lack of data. A kernel density map of sites

(Fig. 1b) suggests the regions for which we have power. We employed a number of methods to control biases and aid interpretation for regions with coverage in our data set: (1) we pruned and merged dates that occur within 200-year bins or within 0.1 degrees latitude or longitude, reducing differences owing to funding or interest of excavation; (2) we compared date and site densities (Fig. 3), as sites are less susceptible to excavation or funding preferences; (3) we spatially smoothed date occurrences using kernel density maps (Fig. 2). Therefore, occurrence of a date in our data set is not interpreted as occupation of the site alone, but rather evidence of human presence in the local region over a 1,000-year period.

Calibration curve effects. Notably, the calibration curve itself can introduce peaks and troughs that look similar to population size changes. We took multiple steps to ensure that our results are robust to the effects of the calibration curve on SCPDs. First, we plotted a 400-year moving average of the density (Fig. 3b), as suggested by ref. 14. Additionally, we plotted histograms of the frequency of occupied sites over time, both calibrated and uncalibrated (Fig. 3c). Importantly, estimates of growth rates and population sizes are based off the best-fit curves, which do not explicitly show calibration effects. We tested 500-year and 1,000-year minimums on phase period for curve fitting, to minimize over-fitting to calibration artefacts. Analyses are focused on the calibrated date range 14,000 to 2,000 calibrated years before present, but we include dates on either side to minimize edge effects.

The South American continent crosses the equator. Therefore, our data set is comprised of dates in both the Southern and Northern Hemispheres. For our analyses, all dates are calibrated using the South Hemisphere curve[32]. A reasonable alternative would be to calibrate dates above the equator using IntCal13 (ref. 33). The SCPD using IntCal13 is highly correlated with our SCPD using only SHCal13 ($R = 0.99$, Pearson correlation coefficient), showing no qualitative differences. This is expected, as calibration differences between the two curves are on the order of 15 years[50], and only 6.2% of dates are above the equator. The division of the two calibration curves by hemisphere is not precise. Circulatory patterns across the South American continent, and therefore ^{14}C levels and calibration effects, are more similar to each other than northern South American patterns are to Northern Hemisphere continents, particularly as all our Northern Hemisphere dates fall within 15 degrees of the equator, where the Hadley cells function[51]. We therefore used a single curve for all dates, SHCal13.

Simulating constant size. As the calibration curve can introduce artificial troughs and peaks into the SCPD, we followed the method of Shennan *et al.*[12] to test if the changes in density from 9 to 5.5 ka are consistent with a constant population size or if there is evidence the population is oscillating around a constant mean. Through simulation of SCPDs, we constructed bounds for the density over time under the null hypothesis of constant size.

For a constant population size, calendar dates (not radiocarbon dates) are expected to follow a Poisson distribution. We consider the date range 9 to 5.5 ka, simulating N dates from a uniform distribution of integers between 10 and 4.5 ka to avoid edge effects. We used $n = 1{,}121$, the number of dates in our pruned data sample with calibrated medians occurring in this time interval. Both samples were taken using the *randsample* command in MATLAB[52]. Using the *R_simulate* function in OxCal, and assigning radiocarbon errors by randomly sampling (with replacement) from errors in the database, we converted the simulated calendar dates to radiocarbon dates. Then, as with the data, we use the sum function in OxCal to create the SCPD for each simulation. Normalizing the SCPD from each simulation and from the data to 1 during the time interval from 9 to 5.5 ka, we considered the distribution of densities at 5-year time intervals. The 98% interval for 200 simulations is considered a guide rather than a distribution from which to calculate *P* values because we did not account for autocorrelation and multiple testing.

Quickly varying from the upper bound to lower bound of the interval for multiple cycles as observed is unlikely and may suggest oscillatory population size. Extended Data Fig. 3 shows a histogram of the variance in normalized density for each of the 200 simulation runs. The observed variance in the SCPD for South America 9 to 5.5 ka is outside the simulated distribution, and is 2.3× the mean of the distribution. The variance is a conservative statistic as it does not account for clustering of densities above or below the simulated confidence interval.

A two-phase model for population size. To determine the general patterns that govern population size over time in South America, we fit 9 models of piecewise exponential and logistic growth (Extended Data Table 1) to the period between 14 to 2 ka. Evidence before ~14.5 ka is still controversial, so we picked this range to avoid biases and edge effects from calibration and sampling. Possible earlier colonization is not excluded based on our analyses, though lack of dense radiocarbon dates suggests any occupation would be at very low densities.

We fit the curves using *nlinfit* in MATLAB, which uses iterative least squares estimation to fit a nonlinear regression to a user-specified function. For exponential growth, we used the two-parameter model $y = p_0 e^{rx}$, where x is time, p_0 is an initial value and r is growth rate. For logistic growth, we have a three-parameter model, $y = \frac{K}{1 + e^{r(x - p_0)}}$, where K is carrying capacity or the limit of the logistic function, r is the steepness of the curves, and p_0 is the midpoint of the sigmoid.

For models with multiple pieces, which we refer to as phases, we add a parameter for the change point from one curve to the next, allowing for discontinuities between neighbouring curves. To prevent over-fitting to calibration artefacts, we require phases to last at least 1,000 years, and consider change points in 500-year intervals.

The SCPD is a convolution of the distribution of calendar dates, and therefore the densities at nearby time periods are not independent. Additionally, the distributions of the calendar dates are nonstandard and vary for each calibrated date. That is, while radiocarbon dates can be represented as a normally distributed random variable, once calibrated, the probability distribution of a date is no longer an exponential family distribution, and differs for each date. Therefore we cannot simply calculate the likelihood of the data under the fit logistic and exponential models. Rather, we use a proxy for the log-likelihood to enable use of likelihood methods for parameter estimation and model choice. Specifically, we took a random sample of dates from the SCPD ($n = 2{,}576$, the number of merged dates), and assumed sample dates are independent ($x_i \in x_1, \ldots, x_{2576}$). Using the models fit to the SCPD, we treated the sampled dates, x_i, as the observed data, calculating a quantity that measures the fit of the model based on model log-likelihood calculations. That is, for exponential and logistic models, respectively, we calculated a proxy for the likelihoods as

$$L(p_0, r; x_1, \ldots, x_n) = \prod_{i=1}^{n} p_0 e^{rx_i},$$

$$L(p_0, r, K; x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{K}{1 + e^{-r(x_i - p_0)}}.$$

The proxy log-likelihoods and associated maximum likelihood estimates of phase-change dates for each mode are presented in Extended Data Table 1. We calculated Bayesian information criteria[53] to determine the best-fit model: a two-phase model with logistic growth from 14 to 5.5 ka and exponential growth from 5.5 to 2 ka. The second best model was of similar form, logistic growth followed by two phases of exponential growth, though it is not well supported by the data, with a difference in BIC from the best-fit model, ΔBIC, of 11.7.

The timing of renewed exponential growth is of demographic and cultural importance. To further examine the estimated change point for two-phase logistic-exponential model, we calculated the proxy likelihood of the model varying the date of the change from logistic to exponential growth in 100-year increments in the range 10 to 2 ka. We found similar support for dates between 5.7 to 4.1 ka ($\Delta BIC < 3$).

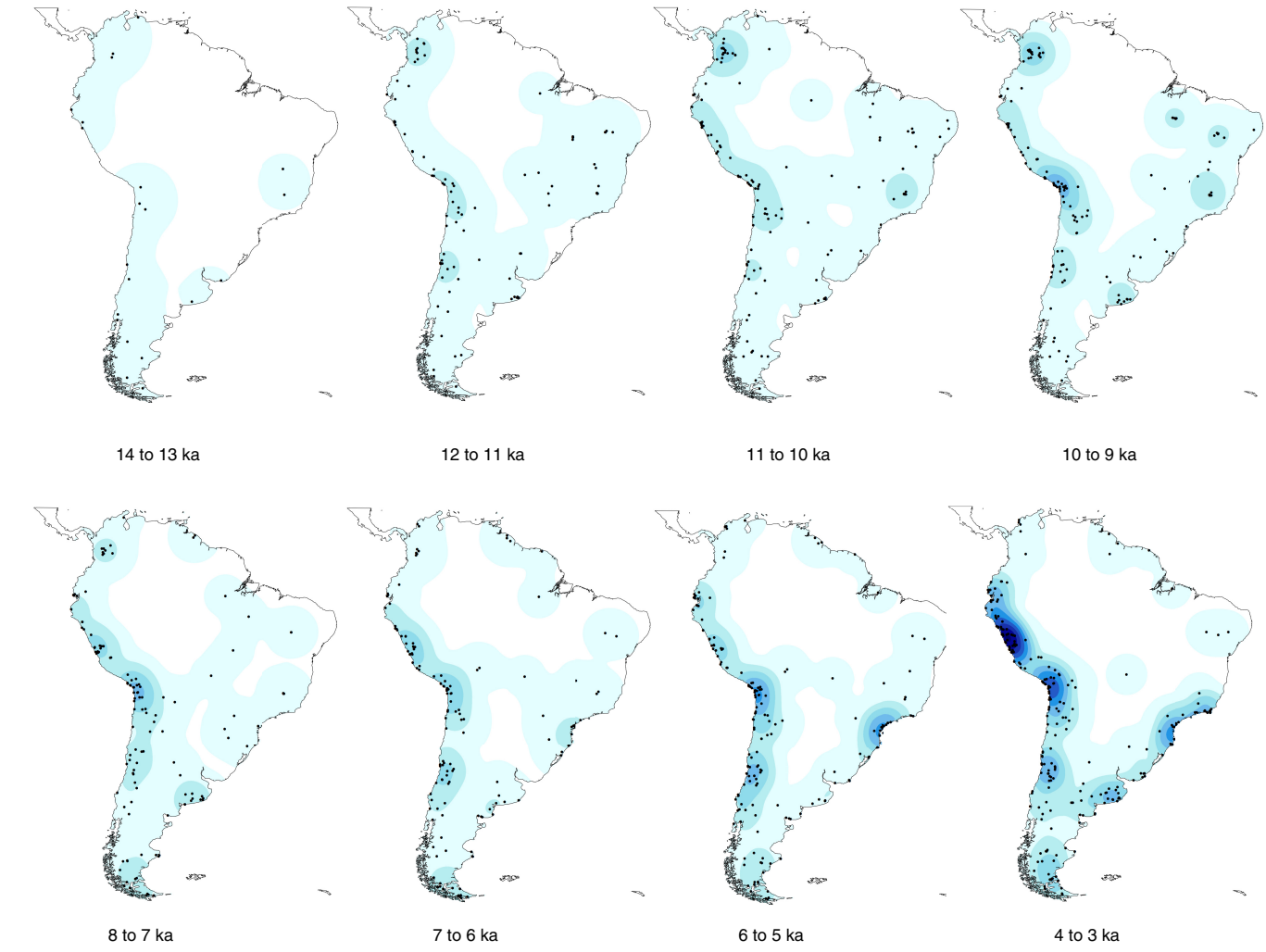
We estimate yearly growth rates over time under the model (Fig. 4a) by calculating $(y_{x+1} / y_x) - 1$. Growth rates should be interpreted as approximate, as they may be a composite of growth and taphonomic processes.

Estimating population size. *Ratio of the density of radiocarbon data over time.* Assuming the SCPD is proportional to population size over time, the ratio of ending to starting density is a measure of the relative increase in population size over that time period. To buffer edge effects, especially as the earliest rare dates are a poor measure of population size, we considered instead the mean density of the 500 years surrounding 13.5 ka to the 500-year mean surrounding 2.5 ka. That is, we calculate the ratio of the mean density from 13.75 to 13.25 ka to the mean density from 2.75 to 2.25 ka. This ratio is 993; interpreted as the population size at 2.5 ka was 993 times larger than that at 13.5 ka.

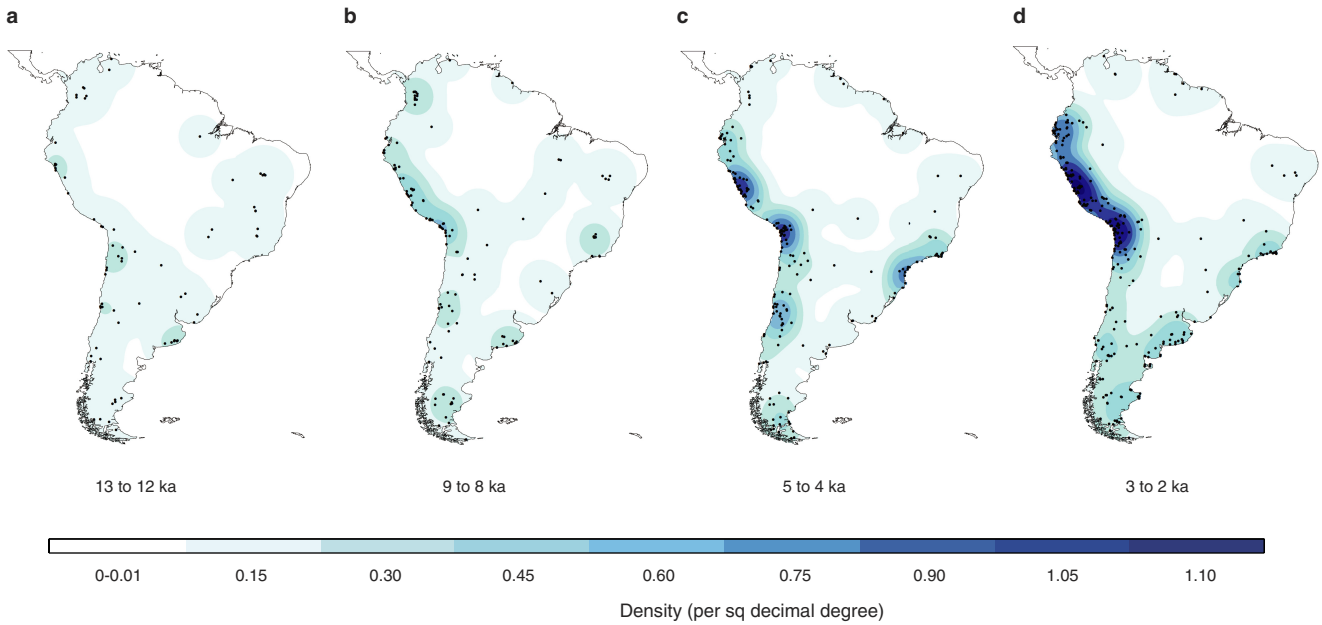
Population size under the two-phase model. As the number of individuals initially entering South America is uncertain, we considered the N-fold growth over time, rather than the census size. We defined N_x as the population size at time x, and have N_x/N_0 as the ratio of population size as time x to initial population size, estimated under the model.

Under the two-phase model, we use the calculated growth rates to iteratively estimate population size over time. We fit the model starting at 14 ka to avoid edge effects of the SCPD, but securely dated archaeological sites place colonization of South America at ~14.8 ka or before[1–5]. Therefore, we extended the fit logistic model to 14.8 ka, starting with a population size $N_0 = 1$ individual, and multiplying the current size by the growth rate at that time. That is, the population size at time x is given by $N_x = N_{x-1}(r_x + 1)$, for growth rate r, which is equal to $N_x = N_{x-1}(y_x / y_{x-1})$. Carrying capacity for South America was reached approximately 8.5 to 9 ka, with $N_x/N_0 = 274$. By 2 ka, $N_x/N_0 = 616$.

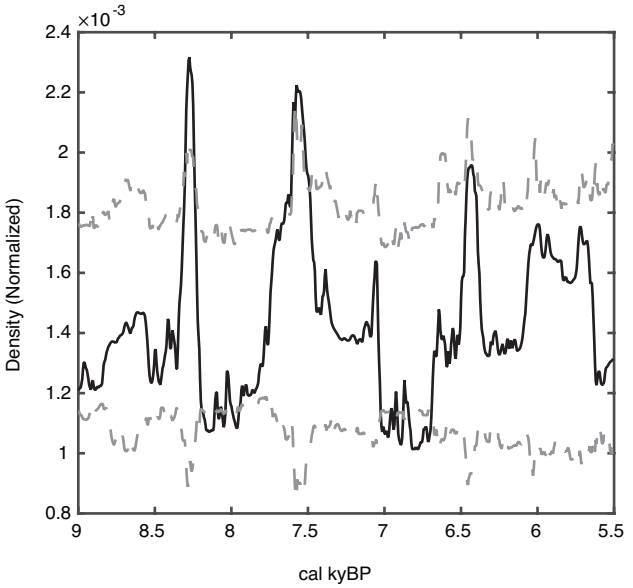
31. Ramsey, C. B. Bayesian analysis of radiocarbon dates. *Radiocarbon* **51,** 337–360 (2009).
32. Hogg, A. G. *et al.* SHCal13 Southern Hemisphere calibration, 0–50,000 years cal bp. *Radiocarbon* **55,** 1889–1903 (2013).
33. Reimer, P. J. *et al.* IntCal13 and Marine13 radiocarbon age calibration curves 0–50,000 years cal bp. *Radiocarbon* **55,** 1869–1887 (2013).
34. Gayo, E. M., Latorre, C. & Santoro, C. M. Timing of occupation and regional settlement patterns revealed by time-series analyses of an archaeological radiocarbon database for the South-Central Andes (16°–25° S). *Quat. Int.* **356,** 4–14 (2015).
35. Thompson, L. G. *et al.* A 25,000 year tropical climate history from Bolivian ice cores. *Science* **282,** 1858–1864 (1998).
36. Petit, J. R. *et al.* Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature* **399,** 429–436 (1999).
37. ArcGIS Version 10 (Redlands CA: Environmental Systems Research Institute Inc., 2010).
38. Silverman, B. W. *Density Estimation for Statistics and Data Analysis* (Chapman & Hall, London, 1986).
39. Epanechnikov, V. A. Non-parametric estimation of a multivariate probability density. *Theory Probab. Appl.* **14,** 153–158 (1969).
40. Wand, M. P. & Jones, M. C. *Kernel Smoothing* (Crc Press, 1994).
41. Chaput, M. A. *et al.* Spatiotemporal distribution of Holocene populations in North America. *Proc. Natl Acad. Sci. USA* **112,** 12127–12132 (2015).
42. Brown, W. A. Through a filter, darkly: population size estimation, systematic error, and random error in radiocarbon-supported demographic temporal frequency analysis. *J. Archaeol. Sci.* **53,** 133–147 (2015).
43. Hinz, M., Feeser, I., Sjögren, K. G. & Müller, J. Demography and the intensity of cultural activities: an evaluation of Funnel Beaker Societies (4200–2800 cal bc). *J. Archaeol. Sci.* **39,** 3331–3340 (2012).
44. Downey, S. S., Bocaege, E., Keri, T. E., Edinborough, K. & Shennan, S. The Neolithic demographic transition in Europe: correlation with juvenility index supports interpretation of the summed calibrated radiocarbon date probability distribution (SCDPD) as a valid demographic proxy. *PLoS ONE* **9,** e105730 (2014).
45. Woodbridge, J. *et al.* The impact of the Neolithic agricultural transition in Britain: a comparison of pollen-based land-cover and archaeological ^{14}C date-inferred population change. *J. Archaeol. Sci.* **51,** 216–224 (2014).
46. Contreras, D. A. & Meadows, J. Summed radiocarbon calibrations as a population proxy: a critical evaluation using a realistic simulation approach. *J. Archaeol. Sci.* **52,** 591–608 (2014).
47. Michczynska, D. & Pazdur, A. Shape analysis of cumulative probability density function of radiocarbon dates set in the study of climate change in the Late Glacial and Holocene. *Radiocarbon* **46,** 733–744 (2004).
48. Michczynska, D. J., Michczynska, A. & Pazdur, A. Frequency distribution of radiocarbon dates as a tool for reconstructing environmental changes. *Radiocarbon* **49,** 799–806 (2007).
49. Surovell, T. A., Finley, J. B., Smith, G. M., Brantingham, P. J. & Kelly, R. Correcting temporal frequency distributions for taphonomic bias. *J. Archaeol. Sci.* **36,** 1715–1724 (2009).
50. Reimer, P. J. *et al.* IntCal13 and Marine13 radiocarbon age calibration curves 0–50,000 years cal BP. *Radiocarbon* **55,** 1869–1887 (2013).
51. Ogburn D. E. Reconceiving the chronology of Inca imperial expansion. *Radiocarbon* **54,** 219–237 (2012).
52. Seidel, D. J., Fu, Q., Randel, W. J. & Reichler, T. J. Widening of the tropical belt in a changing climate. *Nature Geosci.* **1,** 21–24 (2008).
53. MATLAB Release 2015a (The MathWorks, Natick, Massachusetts, USA).
54. Schwarz, G. E. Estimating the dimension of a model. *Ann. Stat.* **6,** 461–464 (1978).



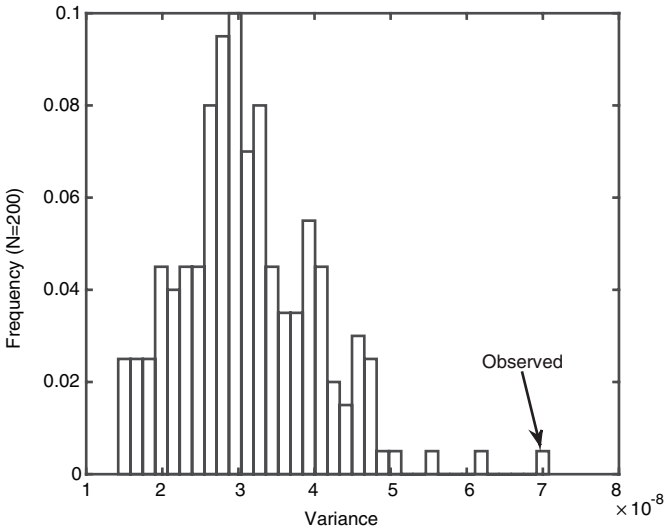
Extended Data Figure 1 | Kernel Density estimates of occupied area for alternative time bins. Estimates of occupied area in 1,000-year time bins for dates not present in Fig. 2, considering uniquely occupied sites in each time bin.



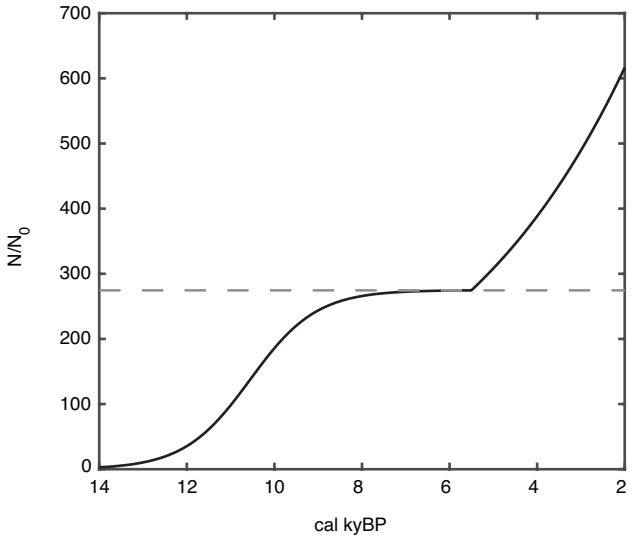
Extended Data Figure 2 | Kernel Density estimates of occupied area for fixed search radius. Estimates of occupied area in 1,000-year time bins for dates present in Fig. 2, using a fixed radius for all time slices of 660 km.



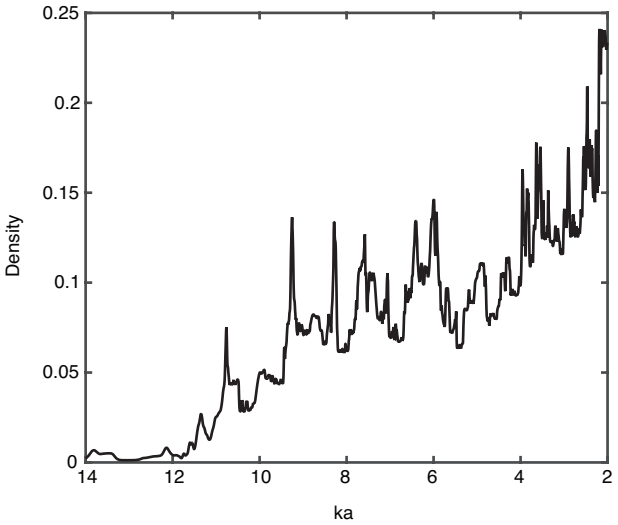
Extended Data Figure 3 | Statistical evaluation of constant population size in the mid-Holocene. The simulated 98% confidence interval (grey) for radiocarbon density under a constant population size in 5-year bins, with the observed SCPD for the data (black).



Extended Data Figure 4 | Observed oscillations are outside the variation of simulated constant size. A histogram of the variance of 200 simulated constant size SCPDs, with the observed variance of density in the observed SCPD from 9 to 5.5 ka outside the distribution of those simulated.



Extended Data Figure 5 | Inferred population size over time. Assuming colonization of South America 14.8 ka by a population of size N_0, we plot the relative increase in population size over time under the two-phase model (black). Carrying capacity (grey) occurs approximately 8.5 ka with a relative population size $N/N_0 = 274.6$, and a final population size at 2 ka of $N/N_0 = 616.6$.



Extended Data Figure 6 | SCPD for Andean/Pacific Coastal Region. The Pacific Coast region and Andes, which represents 52% of dates and multiple of the major cultural centres, show a similar trend as the SCPD for the continent.

Extended Data Table 1 | Model likelihoods and model choice

Model	Transitions (ka)	Log likelihood	BIC	ΔBIC
Single phase, exponential	N/A	-4335.45	8686.56	264.91
2 phases, exponential	10	-4208.63	8448.59	26.93
3 phases, exponential	10; 5.5	-4195.09	8437.18	15.52
4 phases, exponential	11; 7; 5.5	-4190.90	8444.46	22.81
5 phases, exponential	11; 7; 6.5; 5	-4186.96	8452.24	30.59
6 phases, exponential	11; 7; 6.5; 5.5; 4	-4185.06	8464.12	42.47
Single phase, logistic	N/A	-4241.69	8506.88	85.23
2 phases, logistic then exponential	***5.5***	***-4191.25***	***8421.65***	***0***
3 phases, logistic then 2 exponential	4; 3	-4189.27	8433.36	11.71

Nine piece-wise exponential and logistic models considered, with associated log-likelihoods and Bayesian information criteria. Best-fit model is emphasized: logistic growth until 5.5 ka followed by exponential growth until 2 ka.

LETTER

doi:10.1038/nature17158

A neuronal circuit for colour vision based on rod–cone opponency

Maximilian Joesch[1] & Markus Meister[2]

In bright light, cone-photoreceptors are active and colour vision derives from a comparison of signals in cones with different visual pigments. This comparison begins in the retina, where certain retinal ganglion cells have 'colour-opponent' visual responses—excited by light of one colour and suppressed by another colour[1]. In dim light, rod-photoreceptors are active, but colour vision is impossible because they all use the same visual pigment. Instead, the rod signals are thought to splice into retinal circuits at various points, in synergy with the cone signals[2]. Here we report a new circuit for colour vision that challenges these expectations. A genetically identified type of mouse retinal ganglion cell called JAMB (J-RGC)[3], was found to have colour-opponent responses, OFF to ultraviolet (UV) light and ON to green light. Although the mouse retina contains a green-sensitive cone, the ON response instead originates in rods. Rods and cones both contribute to the response over several decades of light intensity. Remarkably, the rod signal in this circuit is antagonistic to that from cones. For rodents, this UV-green channel may play a role in social communication, as suggested by spectral measurements from the environment. In the human retina, all of the components for this circuit exist as well, and its function can explain certain experiences of colour in dim lights, such as a 'blue shift' in twilight. The discovery of this genetically defined pathway will enable new targeted studies of colour processing in the brain.

Like most mammals, the mouse has one type of rod and two types of cone photoreceptors, with absorption maxima in the ultraviolet (S pigment) and green (M pigment) region of the spectrum. As in other small mammals, the retinal organization of the cones is inhomogeneous: the M and S pigments are largely segregated in the dorsal and ventral retina, respectively[4]. At the level of ganglion cells, the spectral sensitivity essentially follows this cone distribution[5,6], which severely limits any local comparison of signals across cone pigments. Because behavioural experiments show that mice can indeed 'see colour'[7], it has been suggested that colour vision in mice operates on very different principles from primates[8]. Surprisingly, as we demonstrate here, the mouse does have a dedicated ganglion cell type with clearly opponent responses to light of different wavelengths. It uses an unexpected retinal circuit that circumvents the obstacle caused by the spatial segregation of cone pigments.

We recorded the visual responses of J-RGCs in the retina of a mouse line that labels these neurons fluorescently[3] (Fig. 1a). When probed with white light, the receptive field has OFF-type sensitivity in the centre and ON-type sensitivity in the surround (Fig. 1b). As reported previously, the surround is stronger on the side of the asymmetric dendritic arbor[3]. Stimulation using coloured lights led to a surprise. Many J-RGCs produce an OFF response to uniform UV light, but an ON response to green light (Fig. 1c). The UV-OFF response arises in the receptive field centre and is driven almost entirely by the S pigment (S-OFF), whereas the Green-ON response derives from the surround from almost pure M pigment (M-ON) (Fig. 1d–f and Extended Data Fig. 1a).

A survey of J-RGCs across the retina revealed diversity in their spectral sensitivities (Fig. 2a, b and Extended Data Fig. 1). The receptive field centre reflects the dorsoventral cone opsin gradient[5] (Fig. 2c), S-OFF ventrally and M-OFF dorsally. The surround, however, had an M-ON spectrum regardless of retinal location (Fig. 2a, b). Focusing

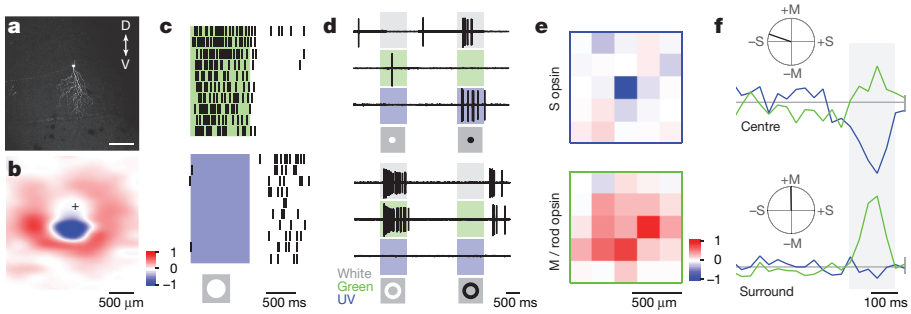


Figure 1 | A spectrally opponent pathway in the mouse retina. **a**, A fluorescent J-RGC in a whole-mounted retina from a JAM-B-CreER; Thy1-STOP-YFP double transgenic mouse. D, dorsal, V, ventral. **b**, Spatial receptive field of a different J-RGC obtained by reverse correlation of the intracellular voltage to an achromatic random flicker stimulus (see Methods). Cross indicates soma position. Polarity: red, ON; blue, OFF. **c**, Raster graph of a spectrally opponent J-RGC response to either full field green (top) or UV (bottom) light stimuli (with green adapting background light). **d**, Response to a flashed spot (top, 250 µm diameter) or annulus (bottom 2,000 µm and 350 µm for outer and inner diameter, respectively) centred on the receptive field using UV, green, or white (UV + green) light. **e**, Spatial receptive field (see Methods) split into contributions from S opsin (top) and M/rod opsin (bottom). Polarity: red, ON; blue, OFF. **f**, Temporal filter (normalized) for the receptive field centre (top, centre pixel in **e**) and surround (bottom, average of the 8 pixels surrounding the centre in **e**) for S opsin and M/rod opsin (blue and green traces, respectively). Graph reports the average opsin activation that occurred as a function of time before a spike. Inset: opsin-space polar graph of the chromatic sensitivity in centre (top) and surround (bottom) calculated from the mean values of the M and S curves in the shaded interval (−198 to −33 ms). For example '+M' indicates a pure M-ON response.

[1]Harvard University, 52 Oxford Street, Cambridge, Massachusetts 02138, USA. [2]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California 91125, USA.

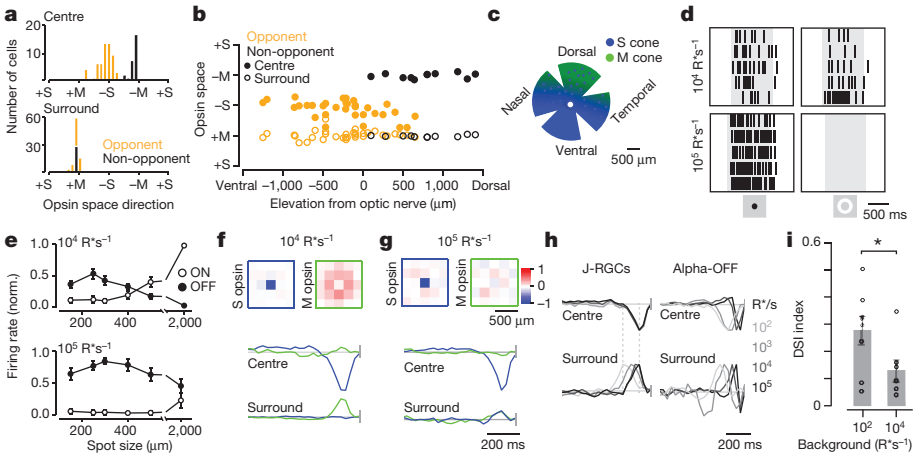


Figure 2 | Rod–cone antagonism. a, Histogram of chromatic sensitivity for centre (top) and surround (bottom) responses of all J-RGCs ($n = 96$). The angle in opsin space is derived as in Fig. 1f. Black, non-opponent cells; orange, opponent cells. **b**, Opsin contributions to the receptive field centre and surround plotted as a function of dorsoventral position. **c**, Cone opsin distribution in a schematic retina drawn with four incisions[5]; (blue, S opsin; green, M opsin; dots, pure S cones). **d**, Response to flashes at intermediate (10^4 R*s^{-1}) and high (10^5 R*s^{-1}) intensity. Achromatic OFF-spots (left, 250 μm diameter) drive the RF centre, and ON-annuli (right, 2,000 μm and 350 μm for outer and inner diameter, respectively) drive the surround. Raster graph of spikes on 5 trials. **e**, Center-surround antagonism for spectrally opponent J-RGCs, at intermediate (top) and high (bottom) intensity. Each curve shows the peak firing rate in response to flashed spots of increasing size, measured separately at light onset (ON) and offset (OFF). Data were normalized for each cell and averaged over 7–12 cells (mean ± s.e.m.). **f**, **g**, Time-course and spatial profile of the receptive field at photopic intensities of 10^4 R*s^{-1} (**f** is normalized to a common peak value and averaged over 67 cells) and 10^5 R*s^{-1} (**g**, 20 cells), displayed as in Fig. 1e, f. **h**, Time-course for centre and surround responses acquired at four mean intensities (10^2, 10^3, 10^4 and 10^5 R*s^{-1}, grey to black), and normalized to a common peak value. Left, J-RGCs (averaged over 8–20 cells); surround kinetics speed up by 125 ms with increasing light intensity (dotted lines), whereas centre kinetics do not. Right, OFF-sustained alpha cells (5 cells). **i**, Direction selectivity of J-RGCs at low and high background intensity. The direction selectivity index (Methods) to small spots moving in eight directions is plotted for each J-RGC, along with mean ± s.e.m. *$P < 0.05$, one-way ANOVA.

attention on the spectrally opponent J-RGCs in the ventral retina, their M-sensitive surround poses a puzzle because the ventral retina is thought to be largely devoid of M-type cones[4]. The only green-sensitive pigment available there in abundance is the rhodopsin in rods. Its absorption spectrum is close to that of the M-cone pigment and indistinguishable by our spectral analysis (Figs 1 and 2 and Extended Data Fig. 2).

To test whether the surround is driven by rods, we systematically varied the absolute light level of the visual stimuli. At light levels that cause isomerization of ~1 rhodopsin per rod per second (1 R*s^{-1}) the sensitivity of rods begins to decline following Weber's law until they cease to function at ~10^5 R*s^{-1} (refs 9 and 10). At intensities of ~10^2 R*s^{-1}, cones are effectively more sensitive than rods, and gradually take over visual signalling[9,11]. As predicted, an annulus flashed on the surround at mean intensities of 10^4 R*s^{-1} still produced ON responses, but at tenfold higher intensity this response was lost (Fig. 2d). By contrast, the centre OFF response strengthened at the highest intensity (Fig. 2d). A weak antagonistic surround remained at the highest light levels, suppressing the OFF response by about 40% from its peak (Fig. 2e). This suppression had a contribution from the S opsin (Fig. 2g), which was not detectable at lower intensities (Fig. 2f). Thus it appears that the antagonistic surround of J-RGCs is mostly driven by rods. Non-opponent J-RGCs in the dorsal retina, however, seem to be less dominated by rods (Extended Data Fig. 3a–c).

At low intensities, the surround response was strong, but considerably slower than that of the centre (Fig. 2h, 10^2 R*s^{-1}). Increasing the light level accelerated the kinetics to approach the centre response (Fig. 2h, 10^5 R*s^{-1}), which retained the same kinetics throughout this intensity range. Because the rod response is considerably slower than that of cones[12], this observation suggests that the surround draws on a rod-driven pathway that gradually saturates at the highest light levels, whereas the centre response is dominated by cones throughout. For comparison we recorded from ON- and OFF-sustained alpha-RGCs, because they can be identified easily by their large soma size and sustained centre response[5,13]. In these RGC types, the kinetics of both centre and surround accelerated with increasing light level, indicating that both receptive field regions receive substantial rod signals (Fig. 2h).

For the J-RGC, the weakening of the surround at high light levels has a strong impact on another functional characteristic—its direction selectivity. This direction preference arises largely from an interaction between the receptive field centre and the asymmetric surround[3]. Consistent with this, we found that adaptation to a background that weakens the surround abolished much of the direction preference to moving spots (Fig. 2i).

To explore the retinal circuits underlying spatial and chromatic opponency in ventral J-RGCs, we voltage-clamped the cells and measured their synaptic currents driven by chromatic stimuli in different parts of the receptive field. The results revealed an even more intricate form of opponency. Both excitatory and inhibitory currents individually have spectrally opponent centre-surround receptive fields. The excitation has a polarity of S-OFF in the centre and M-ON in the surround, whereas inhibition is S-ON in the centre and M-OFF in the surround (Fig. 3a, b and Extended Data Figs 4a and 5). The excitatory current from the receptive field centre most likely reflects glutamate release from type 1 or 2 OFF-bipolar cells[14], whose terminals co-stratify with the dendrites of the J-RGC at the outer margin of the inner plexiform layer[3]. The antagonistic surround of the excitatory current could derive either via direct excitation from ON-bipolar cells that synapse *en passant* onto the J-RGC dendrites or via suppression of the excitatory input from OFF-bipolar cells. The excitatory surround survived pharmacological block of the ON-pathway (Fig. 3d, top right, and 3f) and even simultaneous block of the two inhibitory transmitters GABA and glycine, pointing to the latter hypothesis (Fig. 3e, right, and Extended Data Figs 4c and 5m–o). The only lateral inhibition circuit known to function under such extreme conditions is feedback via horizontal cells to the cones that drive the centre bipolar cells[15] (Fig. 3g).

The inhibitory current driven by the receptive field centre is dependent on GABAergic transmission (Fig. 3c, bottom left, and Extended Data Figs 4b and 5g–l) and thus derives from an ON-amacrine cell type. Block of transmission at the synapse to ON-bipolars eliminated

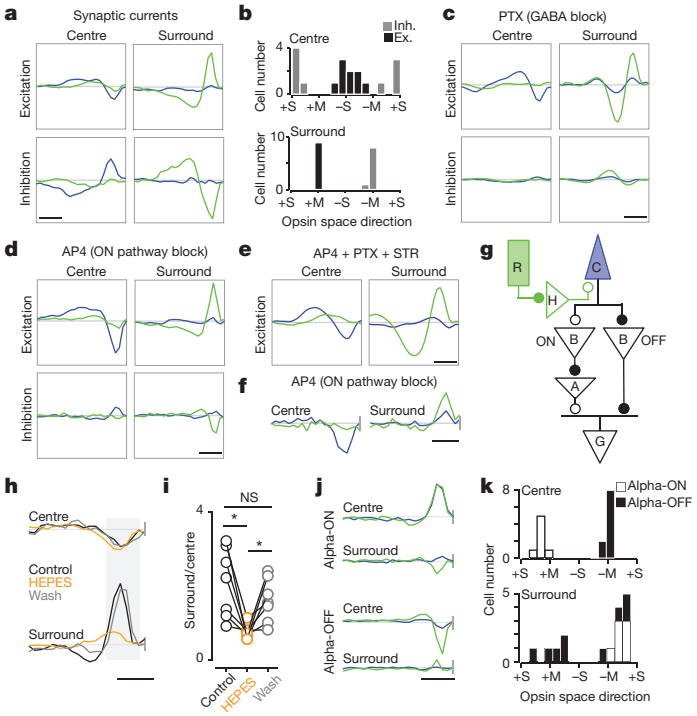


Figure 3 | Synaptic pathways for spectral opponency. a, Visual sensitivity of synaptic currents recorded from J-RGCs under voltage clamp. Excitatory (top) and inhibitory (bottom) conductances are driven by stimulation of the centre (left) or surround (right) of the receptive field. Each curve represents the sensitivity of the conductance to stimulation of the M/rod pigment (green) or S pigment (blue) at various times in the past (normalized to the combined excitation and inhibition, see Methods). Data are from ventral J-RGCs (average of $n = 9$ spectrally opponent cells), where rod and cone signals can be discerned by their spectral sensitivity, but the analysis of dorsal non-opponent J-RGCs yielded similar conclusions (Extended Data Fig. 3d, e). **b**, Histogram of the chromatic sensitivity for centre (top) and surround (bottom) currents (black, excitation; grey, inhibition) recorded from many J-RGCs, displayed by the angle in opsin space, as in Fig. 2a. **c**, As in **a**, but during block of GABA receptors with picrotoxin (PTX; 100 μM; $n = 7$). Although the loss of GABAergic inhibition alters the kinetics of excitatory signals, the centre-surround antagonism remains. **d**, As in **a**, but during block of ON-bipolar responses with L-AP4 (11 μM; $n = 11$). **e**, As in **a**, but during combined block of GABA and glycine receptors and ON-bipolar pathways (100 μM PTX, 10 μM strychnine, 11 μM L-AP4; $n = 7$). **f**, Spiking responses, displayed as in Fig. 1f ($n = 8$) during L-AP4 application. **g**, Working model of neural circuits underlying the response of J-RGCs (G), involving rods (R), cones (C), horizontal cells (H), bipolar cells (B), and amacrine cells (A). Open/closed circles denote inhibitory/excitatory synapses. **h**, Visual sensitivity of spiking responses (average of $n = 7$) in control solution (black), 20 mM HEPES (orange), and after wash-out (grey). **i**, Change in the centre-surround ratio from HEPES exposure calculated from the mean values in the shaded interval in **h** (−198 to −33 ms). **j**, Visual sensitivity of spiking responses from sustained alpha ON−RGCs (average of $n = 7$) and sustained alpha OFF−RGCs (average of $n = 11$). Sensitivity plotted as in Fig. 2f for stimulation of S opsin (blue) and M/rod opsin (green). **k**, Histogram of the chromatic sensitivity for centre (top) and surround (bottom) recorded from alpha ON−RGCs and alpha OFF−RGCs. Display as in Fig. 2a. Scale bars, 200 ms.

~80% of the inhibition (Fig. 3d, bottom left, and Extended Data Fig. 5a–f), indicating that the receptive field centre drives inhibition through ON-bipolars via GABAergic ON-amacrines (Fig. 3g). Again, these inhibitory currents showed an antagonistic OFF-surround, and those responses had the same pharmacological sensitivity as the centre signals (Fig. 3c, bottom right and 3d, bottom right), requiring both the ON-pathway and GABAergic transmission. This suggests that centre and surround use the same bipolar cell pathways and the antagonistic surround already arises in the outer retina as suggested above. We propose that the receptive field centre is implemented by a cone-selective push-pull system of OFF-bipolars and narrow-field ON-amacrines. The surround derives from lateral inhibition of the cones via horizontal cells with rod input (Fig. 3g).

The horizontal cell of the mouse retina connects to cones at the dendritic tree near the soma and to rods at the terminal arborization of its axon. It has been claimed that the thin axon and a high somatic membrane conductance preclude any flow of signals between the soma and terminal compartments[16]. However, transmission from soma to terminal[17] and in the opposite direction has been observed directly[18]. To specifically test the role of horizontal cells we blocked their feedback by applying the pH buffer HEPES[19]. This strongly reduced the surround, on average fourfold relative to the centre response (Fig. 3h, i). At rod-saturating light levels, the horizontal cell should still implement lateral inhibition among cones, and indeed the surround has a UV-sensitive component under those conditions (Fig. 2g). One could imagine additional surround circuits in the inner retina that make use of the ON-type rod bipolar cell, but we found that J-RGCs remain spectrally opponent even under block of all ON-bipolar responses (Fig. 3f).

If the rod-driven surround does arise already at the cone terminal, then one would expect to see the effects in other ganglion cell types. Most of the sustained alpha cells we recorded in the ventral retina indeed showed a surround dominated by the rod pigment (Fig. 3j, k). Unique to the J-RGCs is a centre pathway that is perfectly selective for cones, even under conditions where the rods are clearly active. To test this notion under extreme conditions, we recorded from J-RGCs in a mutant background (*Gnat2*^{cpfl3}) where the cones are silent owing to a transducin mutation. We found that the receptive field centre of J-RGCs had little or no detectable light sensitivity, whereas the surround responded strongly to rod signals (Extended Data Fig. 6). Thus, we conclude that spectral opponency in the mouse retina arises already in certain bipolar cells, by virtue of their cone-selective inputs[20].

The mesopic range of luminance, in which both rods and cones are signalling, spans 0.02–20 cd m^{-2}, namely from dim moonlight to bright twilight[21]. Mice are active under all of these conditions. Although the mouse is commonly considered as strictly nocturnal, in the wild they tend to seek food during the day[22]. A further constraint on this colour

channel is that it can function only in the part of the retina containing S pigment. In a mouse with a head posture typical during locomotion, that part covers the entire superior visual field, dipping below the horizon[23] to include nearby points on the ground.

To explore what ecological benefits mice might draw from colour vision, we searched for objects in the natural world that would stand out in this spectrally opponent channel. Using a modified camera that approximates the absorption of S and M rod pigments, we screened for salient UV-green coloured objects under natural light. Certain seeds and urine marks stand out by being relatively bright or dark in UV[24], respectively (Extended Data Fig. 7). A spot of dry mouse urine is barely detectable from the background with the human eye, but has fourfold higher contrast (Extended Data Fig. 8) when processed by a spectrally opponent system like the J-RGC. Beyond its obvious excretory role, urine serves an important function for social communication among mice. Males mark their territories by squirting urine in characteristic patterns; these marks are sampled and counter-tagged by other individuals, and communicate information regarding social status. Interpretation of the tag requires physical contact because the relevant pheromones are non-volatile[25]. In the wild, mice tag their territory with accumulation of urine and solid matter that lead to sizable vertical structures called urine posts[26]. As expected, this material is dark in the UV (Extended Data Fig. 7d, e). Mice use primarily visual cues to recognize their territory boundaries[27]. On that background, we propose that mice in the wild can identify urine tags visually, using the UV-green opponent colour channel of the retina, which assists in approaching the tag.

In the human retina, the type-1 horizontal cell offers a similar route for opponency between rods and cones. Interestingly, it predominantly contacts the L and M cones over S cones[28]. Therefore, when rods are active in dim light, L and M activity should be suppressed, shifting the spectral balance towards signals from S cones. This may explain the pronounced blue shift we experience in twilight and night-time scenes, and the general observation that 'rod-colour is blue'[29]. Finally, certain individuals lack L and M cones entirely. Nonetheless they report percepts of colour, and psychophysical experiments confirm that they experience a two-dimensional colour space, spanned by the signals from rods and blue cones[30]. We suggest that these 'blue-cone monochromats' use the same rod-cone circuits as the mouse J-RGC for opponent-colour processing.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 29 March 2015; accepted 21 January 2016.
Published online 6 April 2016.

1. Solomon, S. G. & Lennie, P. The machinery of colour vision. *Nature Rev. Neurosci.* **8,** 276–286 (2007).
2. Völgyi, B., Deans, M., Paul, D. & Bloomfield, S. Convergence and segregation of the multiple rod pathways in mammalian retina. *J. Neurosci.* **24,** 11182–11192 (2004).
3. Kim, I.-J. J., Zhang, Y., Yamagata, M., Meister, M. & Sanes, J. R. Molecular identification of a retinal cell type that responds to upward motion. *Nature* **452,** 478–482 (2008).
4. Baden, T. *et al.* A tale of two retinal domains: near-optimal sampling of achromatic contrasts in natural scenes through asymmetric photoreceptor distribution. *Neuron* **80,** 1206–1217 (2013).
5. Chang, L., Breuninger, T. & Euler, T. Chromatic coding from cone-type unselective circuits in the mouse retina. *Neuron* **77,** 559–571 (2013).
6. Wang, Y. V., Weick, M. & Demb, J. B. Spectral and temporal sensitivity of cone-mediated responses in mouse retinal ganglion cells. *J. Neurosci.* **31,** 7670–7681 (2011).
7. Jacobs, G. H., Williams, G. A. & Fenwick, J. A. Influence of cone pigment coexpression on spectral sensitivity and color vision in the mouse. *Vision Res.* **44,** 1615–1622 (2004).
8. Jacobs, G. H., Williams, G. A., Cahill, H. & Nathans, J. Emergence of novel color vision in mice engineered to express a human cone photopigment. *Science* **315,** 1723–1725 (2007); comment **318,** 196 (2007).
9. Naarendorp, F. *et al.* Dark light, rod saturation, and the absolute and incremental sensitivity of mouse cone vision. *J. Neurosci.* **30,** 12495–12507 (2010).
10. Soucy, E., Wang, Y., Nirenberg, S., Nathans, J. & Meister, M. A novel signaling pathway from rod photoreceptors to ganglion cells in mammalian retina. *Neuron* **21,** 481–493 (1998).
11. Williams, G. A., Daigle, K. A. & Jacobs, G. H. Rod and cone function in coneless mice. *Vis. Neurosci.* **22,** 807–816 (2005).
12. Nikonov, S. S., Kholodenko, R., Lem, J. & Pugh, E. N. Jr Physiological features of the S- and M-cone photoreceptors of wild-type mice from single-cell recordings. *J. Gen. Physiol.* **127,** 359–374 (2006).
13. Pang, J.-J., Gao, F. & Wu, S. M. Light-evoked excitatory and inhibitory synaptic inputs to ON and OFF alpha ganglion cells in the mouse retina. *J. Neurosci.* **23,** 6063–6073 (2003).
14. Euler, T., Haverkamp, S., Schubert, T. & Baden, T. Retinal bipolar cells: elementary building blocks of vision. *Nature Rev. Neurosci.* **15,** 507–519 (2014).
15. Thoreson, W. B. & Mangel, S. C. Lateral interactions in the outer retina. *Prog. Retin. Eye Res.* **31,** 407–441 (2012).
16. Nelson, R., von Litzow, A., Kolb, H. & Gouras, P. Horizontal cells in cat retina with independent dendritic systems. *Science* **189,** 137–139 (1975).
17. Trümpler, J. *et al.* Rod and cone contributions to horizontal cell light responses in the mouse retina. *J. Neurosci.* **28,** 6818–6825 (2008).
18. Szikra, T. *et al.* Rods in daylight act as relay cells for cone-driven horizontal cell-mediated surround inhibition. *Nature Neurosci.* **17,** 1728–1735 (2014).
19. Hirasawa, H. & Kaneko, A. pH changes in the invaginating synaptic cleft mediate feedback from horizontal cells to cone photoreceptors by modulating Ca^{2+} channels. *J. Gen. Physiol.* **122,** 657–671 (2003).
20. Tsukamoto, Y., Morigiwa, K., Ueda, M. & Sterling, P. Microcircuits for night vision in mouse retina. *J. Neurosci.* **21,** 8616–8623 (2001).
21. Nathan, J. *et al.* Scotopic and photopic visual thresholds and spatial and temporal discrimination evaluated by behavior of mice in a water maze. *Photochem. Photobiol.* **82,** 1489–1494 (2006).
22. Daan, S. *et al.* Lab mice in the field: unorthodox daily activity and effects of a dysfunctional circadian clock allele. *J. Biol. Rhythms* **26,** 118–129 (2011).
23. Sterratt, D. C., Lyngholm, D., Willshaw, D. J. & Thompson, I. D. Standard anatomical and visual space for the mouse retina: computational reconstruction and transformation of flattened retinae with the Retistruct package. *PLoS Comput. Biol.* **9,** e1002921 (2013).
24. Tovée, M. J. Ultra-violet photoreceptors in the animal kingdom: their distribution and function. *Trends Ecol. Evol.* **10,** 455–460 (1995).
25. Hurst, J. L. & Beynon, R. J. Scent wars: the chemobiology of competitive signalling in mice. *Bioessays* **26,** 1288–1298 (2004).
26. Welch, J. F. Formation of urinating posts by house mice (*Mus*) held under restricted conditions. *J. Mamm.* **34,** 502–503 (1953).
27. Mackintosh, J. H. Factors affecting recognition of territory boundaries by mice (*Mus musculus*). *Anim. Behav.* **21,** 464–470 (1973).
28. Ahnelt, P. & Kolb, H. Horizontal cells and cone photoreceptors in human retina: a Golgi-electron microscopic study of spectral connectivity. *J. Comp. Neurol.* **343,** 406–427 (1994).
29. Stabell, U. & Stabell, B. Mechanisms of chromatic rod vision in scotopic illumination. *Vision Res.* **34,** 1019–1027 (1994).
30. Reitner, A. & Sharpe, L. T. Is colour vision possible with only rods and blue-sensitive cones. *Nature* **352,** 798–800 (1991).

Acknowledgements We thank E. Soucy and J. Greenwood for technical support, J. Cauceglia for providing the urine post samples, J. R. Sanes and E. Soucy for comments on the manuscript. This work was supported by grants to M.M. from the NIH and to M.J. from The International Human Frontier Science Program Organization.

Author Contributions M.J. designed the study, performed all experiments, interpreted results, and wrote the manuscript. M.M. helped design the study, interpret results, and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.M. (meister@caltech.edu) or M.J. (joeschkrotki@fas.harvard.edu).

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Histology. For the image in Fig. 1a, the retina was drop-fixed with 4% paraformaldehyde in PBS for 2 h on ice, washed with PBS and blocked with 3% goat serum/0.1% Triton X-100/PBS overnight. For staining, tissue was incubated with 3% goat serum/0.1% Triton X-100/PBS and rabbit anti-GFP Alexa Fluor 488 conjugate (dilution 1:1,000, Invitrogen, A-21311) at 4°C for 3 days and washed with PBS. Retina was mounted on Vectashield mounting medium (Vectorlabs) and imaged in a confocal microscope (Olympus FVA).

Animals. Animals were used in accordance with NIH guidelines and protocols approved by Institutional Animal Use and Care Committee at Harvard University. Mice were maintained on a C57/B6J background. Both male and female mice were used in this study. Animals were 40 to 150 days old at the time of euthanasia. To visualize J-RGCs, JAMB-Cre-ER mice were mated to Thy1-STOP-YFP transgenic mice[3]. To activate Cre and thereby YFP, 3 mg of tamoxifen (dissolved in sunflower oil at 50 mM; W530285 & T5648; Sigma) was injected at least twice intraperitoneally at between P10 and P21. Gnat2^{cpfl3} mice[31] were obtained from Jackson Laboratories and crossed to JAMB-Cre-ER and Thy1-STOP-YFP.

Recording. The dark-adapted mouse retina was isolated under far-red light (LED peak 735 nm, additionally filtered with a 735 nm LP filter eliciting an isomerization rate of ~17 R*/s) in oxygenated Ames' medium (Sigma) with constant bubbling (95% O_2, 5% CO_2) at room temperature. Four incisions were made to flat-mount the retina with ganglion cells facing up in a superfusion chamber on the stage of a custom-built upright fluorescence microscope. Ganglion cell bodies were visualized using oblique angled IR light (850 nm LED, eliciting isomerization rates <0.01 R*s^{-1}). Spike recordings were obtained with loose cell-attached patch microelectrodes filled with Ames' medium. Current and voltage recordings were made in whole-cell voltage- and current-clamp modes, respectively (Axon Multiclamp 700B). Electrodes had an impedance of 5–8 MΩ and were filled with high potassium internal solution (120 mM KAc, 1 mM NaCl, 0.2 mM $CaCl_2$, 1 mM $MgCl_2$, 10 mM EGTA, 2 mM MgATP, 0.3 mM GTP, 1 mM KCl, 10 mM HEPES) containing an additional 5 mM lidocaine *N*-ethyl bromide (QX314-Br). In pharmacological experiments, agents were bath-applied at the following concentrations: 100 μM picrotoxin, 10 μM strychnine, and 11 μM L-AP4. All chemicals were obtained from Sigma-Aldrich, with exception of QX-314 (EMD Millipore Biosciences). During voltage-clamp recordings, excitatory and inhibitory synaptic currents were separated by voltage clamping the cell to the equilibrium potential of chloride (−65 mV) and unselective cation channels (0 mV), respectively. A junction potential of 12 mV was subtracted. It is possible that conductances in distal dendrites of the J-RGC are incompletely clamped under this protocol, leading to some uncertainty about the assignment of inhibitory and excitatory inputs. However, such errors are likely small, because PTX blocked almost all the inhibitory currents and not the excitatory currents (Fig. 3c). Only cells with an input resistance of 8–25 MΩ were used. The superfusion liquid was heated to 32 °C (Warner TC-324B). Signals were digitized at 10 kHz (National Instruments PCIe-6321), highpass-filtered at 0.1 or 1 Hz, and acquired using software written in LabVIEW (National Instruments). Data were analysed using MATLAB (MathWorks). Fluorescent J-RGCs were detected by brief excitation (64 ms) with a green LED eliciting ~10^6 R*/flash. This flash was followed by up to 20 min of recovery and adaptation to the intended mean luminance level before recordings were initiated. This restored the RGC response properties to the pre-flash condition (Extended Data Fig. 9). After the physiological recordings, the cell's arbor was revealed with longer fluorescence exposures. To determine a neuron's dorsoventral location on the retina, we measured the distance from the soma to the optic nerve and used the dendritic orientation to identify the ventral direction.

Stimulation. Light stimuli were delivered from a modified Dell M109S DLP projector through a custom-made lens system and focused onto the photoreceptors (frame rate 60 Hz, magnification 5.5 μm per pixel, maximal Michelson contrast: 0.995). The projector's blue LED was replaced with a high-power UV-LED (ProLight 1W UV LED, peak 405 nm), to improve the differential stimulation of S and M pigments. Owing to peculiarities of the projector, the green light includes a small component from the UV LED and vice versa (Extended Data Fig. 2). The relative intensities of the green and UV lights were chosen such that the average output of the stimulator matches the spectrum of daylight (Extended Data Fig. 2); this ensures that the retina is in a realistic state of chromatic adaptation. Intensities and spectra were measured using a calibrated spectrometer (Thorlabs CCS-100) and a digital power meter (Thorlabs S130C sensor). Most experiments were performed at a mean photopic intensity of 10^4 R*s^{-1} per rod. When stated otherwise, the light intensity was changed uniformly by exchanging reflective neutral density filters (Thorlabs) in the light path. The respective cone pigment isomerization rates for all light levels used in this study are depicted in Extended Data Fig. 2. Isomerization rates were determined using opsin templates[32] and assuming that the mouse rod has an optical density at peak absorption wavelength of 0.015 μm^{-1}, a length of 24 μm, a diameter of 1.4 μm and a quantum efficiency of 0.67 (refs 33, 34).

A spatiotemporal white-noise stimulus was presented using a binary pseudorandom sequence, in which the two primary lights (green and UV) varied independently. For checkerboard stimuli (Figs 1e, f and 2f, g and Extended Data Figs. 3b, c and 5), the checker size was 220×220 μm^2. For the achromatic white-noise stimulus, checker dimensions were 60×60 μm^2, and the green and UV lights were flickered synchronously (Fig. 1b). For the spot-annulus flicker stimuli, the centre spot had a diameter of 250 μm, and the surround annulus had inner and outer diameters of 350 μm and 2,000 μm, respectively; spot and annulus were separated by a constant grey annulus (Fig. 3 and Extended Data Figs 3d, e and 4). All white-noise stimuli were presented at 30 Hz update rate. Spot-annulus flashes were presented with the same stimulus dimensions. For coloured flash experiments, either the UV or the green light were changed from an average white background (both lights at half intensity, Fig. 1d and Extended Data Figs 1 and 4a). In achromatic stimuli, both lights were changed together (Fig. 2d and Extended Data Fig 4b, c). Full-field chromatic stimuli were presented on a green adapting background to reduce M opsin excitation from UV light via the β-band; UV spot and annulus flashes were presented without background (Fig. 1c). Moving spot stimuli consisted of a white spot (width 250 μm, 2×10^4 R*s^{-1}) on a grey background (either 50 R*s^{-1} or 10^4 R*s^{-1}) moved through the receptive field centre in eight different directions chosen randomly and repeated 3 times per experiment with a 1 s pause between sweeps at 800 μm s^{-1} (Fig. 2i).

Analysis. The response to flashing spots or annuli was quantified by counting spikes in the interval between 0.1 s and 1 s after the onset or the offset of the flash, and normalizing to the maximum value obtained. The resulting relative response strength was then averaged across cells (Fig. 2e and Extended Data Fig. 3a). In voltage clamp experiments (Extended Data Fig. 4), the difference between the peak of the stimulus-evoked current and the mean value in the 500 ms interval prior the stimulus onset was analysed.

To measure the response to moving spots, we averaged the spike count over several trials with identical sweeps of the spot. From experiments with eight directions (separated by 45°), the direction selectivity index (DSI) was calculated from the response-weighted vector sum of all directions:

$$\mathrm{DSI} = \frac{|\sum_k r(\varphi_k) e^{i\varphi_k}|}{\sum_k r(\varphi_k)} \quad (1)$$

Where φ_k is the angle of the kth direction and $r(\varphi_k)$ is the corresponding spike rate. This index ranges from 0 for a cell with equal responses to all directions to 1 for a cell that responds to only one direction (Fig. 2i).

The spatiotemporal receptive fields were computed starting with reverse-correlation functions to the randomly flickering stimulus:

$$h(\mathbf{x}, t) = \frac{1}{T}\int_0^T r(t')s(\mathbf{x}, t'+t)dt' \quad (2)$$

where T represents the duration of the recording; $s(\mathbf{x}, t)$ represents the stimulus at location $\mathbf{x}$ and time t; and $r(t)$ represents the response at time t.

The response variable $r(t)$ is either the firing rate, namely the number of spikes per stimulus update interval (Figs 1e, f, 2f–h and 3f, h–j and Extended Data Figs 3b, c and 9), or the membrane potential (Fig. 1b), or membrane currents (Fig. 3a, c–e and Extended Data Figs 3d and 5). The stimulus variables are the normalized intensity of the green or UV lights:

$$s_l = \frac{I_l - \bar{I}_l}{\bar{I}_l}, l = G(\text{green}) \text{ or } U(\text{UV}) \quad (3)$$

where

I_l = intensity of light l
$\bar{I}_l$ = average intensity of light l

From these two correlation functions we derived the sensitivity of the response to modulation of the S and M pigments. Note that the mouse has 3 photoreceptor types, but the rod and M cone have very similar absorption spectra (peak wavelength $\lambda_{max} = 502$ nm versus 508 nm) that cannot be resolved by our methods. For simplicity, we refer to this common spectral sensitivity as 'M opsin', bearing in mind that any response components with that spectral sensitivity may derive from the M cone or from the rod. We arrive at that distinction by independent methods, as described in the text.

Each of the two lights of the stimulator drives both pigments, though in different ratios. To make the conversion, we model the response as

$$r = r_0 + c_S \frac{I_S - \bar{I}_S}{\bar{I}_S} + c_M \frac{I_M - \bar{I}_M}{\bar{I}_M} \quad (4)$$

where

r_0 = baseline response

$\frac{I_p - \bar{I}_p}{\bar{I}_p}$ = Weber contrast for photoreceptor type p, $p = M$ or S

c_p = sensitivity to photoreceptor type p

By evaluating the reverse-correlation to the two lights, h_U and h_G, predicted from this linear model one finds the relation

$$\begin{aligned} h_U &= c_S E_{SU} + c_M E_{MU} \\ h_G &= c_S E_{SG} + c_M E_{MG} \end{aligned} \quad (5)$$

where

$$E_{pl} = \frac{e_{pl}}{e_{pU} + e_{pG}}$$

e_{pl} = isomerization rate in photoreceptor p under average light l

Finally, the respective photoreceptor contributions to the response are

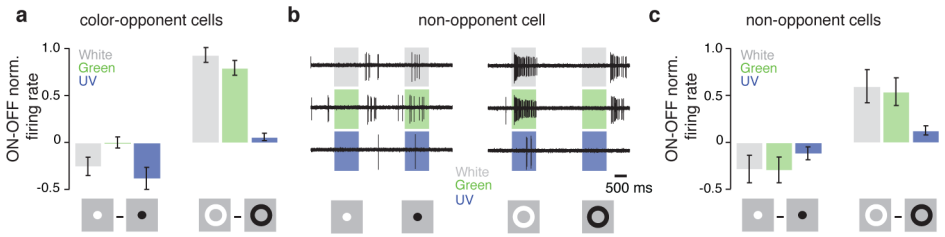
$$\begin{pmatrix} c_S \\ c_M \end{pmatrix} = \mathbf{E}^{-1} \cdot \begin{pmatrix} h_U \\ h_G \end{pmatrix}, \text{ where } \mathbf{E} = \begin{pmatrix} E_{SU} & E_{MU} \\ E_{SG} & E_{MG} \end{pmatrix} \quad (6)$$

In general, these opsin contributions to the response were computed for every location and time point, yielding $c_S(\mathbf{x}, t)$ and $c_M(\mathbf{x},t)$ (as in Fig. 1e). For a specific location and time (for example, Fig. 1f) we represented the ratio of the two opsin contributions by the 'opsin space angle'

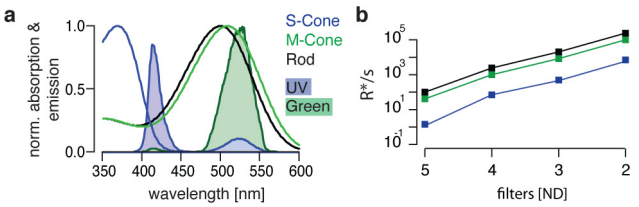
$$\varphi = \arg(c_S + i \cdot c_M) \quad (7)$$

Camera and spectra. We assembled the UV/Green imaging device using a monochromatic camera (Point Grey, FL3-U3-13Y3M-C) with a CMOS chip sensitive over the band 350–950 nm (min 30% of peak sensitivity). We used fused silica lenses (Thorlabs) that are transparent over 200–1,200 nm. Light was filtered using two different band-pass filters that overlap to ~70% with the mouse M and S bands (BP390 #86-348 and BP510 #84-097, Edmund Optics). Images were taken separately in the two bands and adjusted to compensate for the camera's sensitivity to different wavelengths and for chromatic aberration. The illumination was indirect sunlight (Extended Data Fig. 7). Spectra were acquired using a calibrated spectrometer (Thorlabs CCS-100). The spectra in Extended Data Fig. 7f were calculated using opsin templates[32] and multiplied by the transmission spectrum of the mouse eye[35].

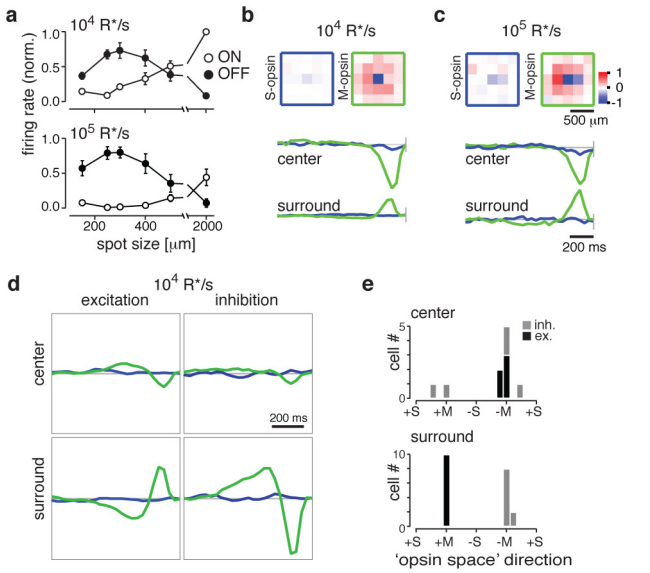
31. Chang, B. *et al.* Cone photoreceptor function loss-3, a novel mouse model of achromatopsia due to a mutation in *Gnat2*. *Invest. Ophthalmol. Vis. Sci.* **47,** 5017–5021 (2006).
32. Govardovskii, V. I., Fyhrquist, N., Reuter, T., Kuzmin, D. G. & Donner, K. In search of the visual pigment template. *Vis. Neurosci.* **17,** 509–528 (2000).
33. Carter-Dawson, L. D. & LaVail, M. M. Rods and cones in the mouse retina. I. Structural analysis using light and electron microscopy. *J. Comp. Neurol.* **188,** 245–262 (1979).
34. Penn, J. S. & Williams, T. P. A new microspectrophotometric method for measuring absorbance of rat photoreceptors. *Vision Res.* **24,** 1673–1676 (1984).
35. Henriksson, J. T., Bergmanson, J. P. & Walsh, J. E. Ultraviolet radiation transmittance of the mouse eye and its individual media components. *Exp. Eye Res.* **90,** 382–387 (2010).



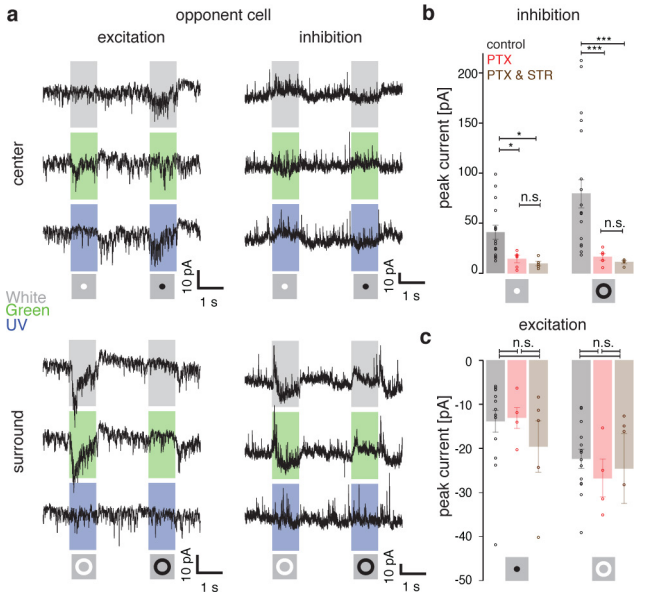
Extended Data Figure 1 | Spiking responses to chromatic centre and surround stimuli. a, Summary of responses to centre (left) and surround (right), derived from the experiment of Fig. 1d. Peak firing rate to an ON flash was subtracted from that to the OFF flash, and averaged over cells ($n = 7$; mean ± s.e.m.). Green light acts almost exclusively in the surround, UV light only in the centre. **b**, Response of a non-opponent J-RGC to a flashed spot or annulus (as in Fig. 1c). **c**, Summary of the differential response for non-opponent cells ($n = 8$, ± s.e.m.) displayed as in **a**. Green light and UV light act with the same polarity, OFF in centre, ON in surround. Note that both lights excite the M cone that is prevalent in the dorsal retina.



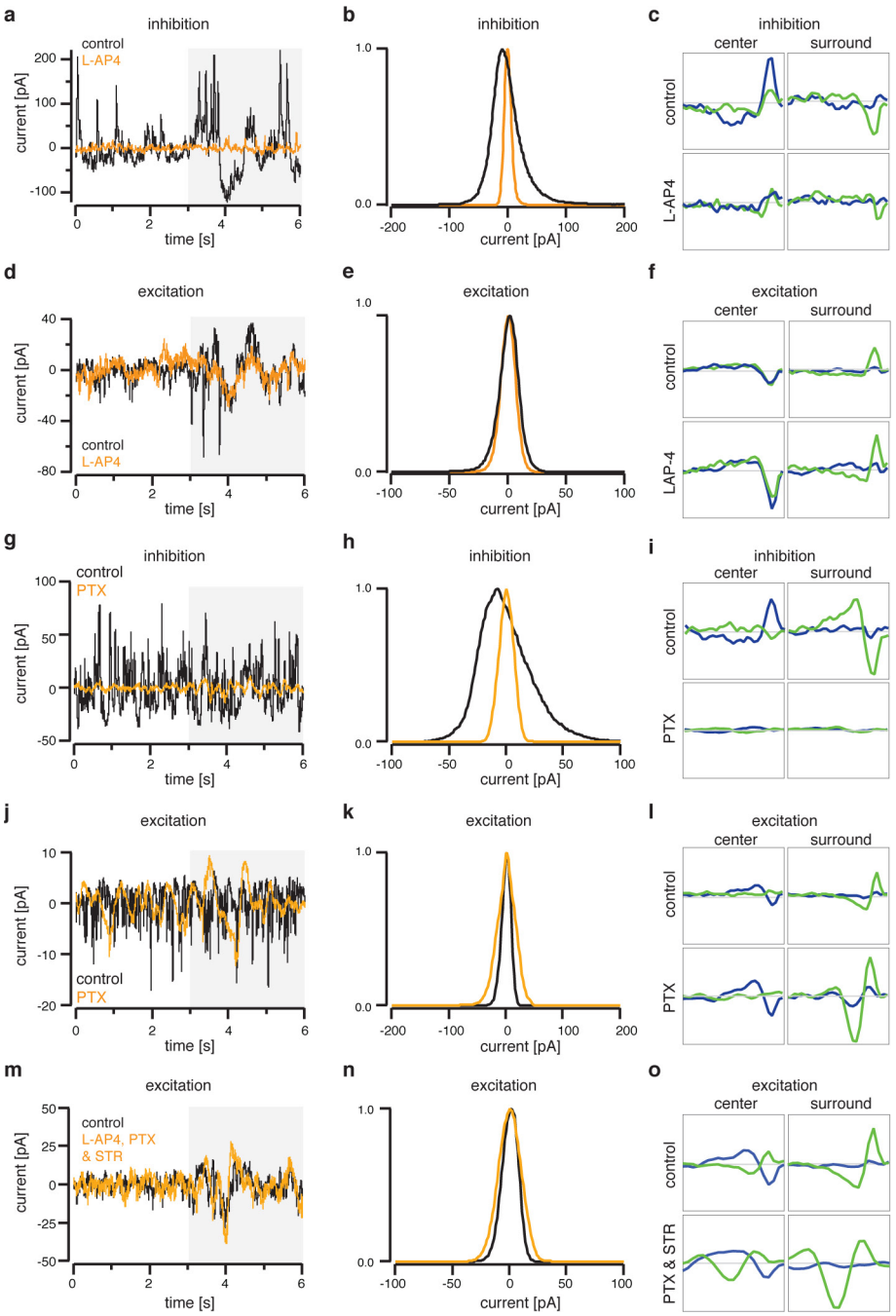
Extended Data Figure 2 | Spectra of the visual stimuli. a, Normalized absorption spectra of mouse photoreceptors (black trace: rhodopsin; green: M opsin; blue: S opsin). Overlaid, the normalized emission spectra of the UV and green light emitted by the DLP projector (filled blue: UV, filled green: green light). **b**, Isomerization rate per photoreceptor in rods (black), M cones (green) and S cones (blue). The collecting area for cones was 0.2 μm^2, for rods see Methods.



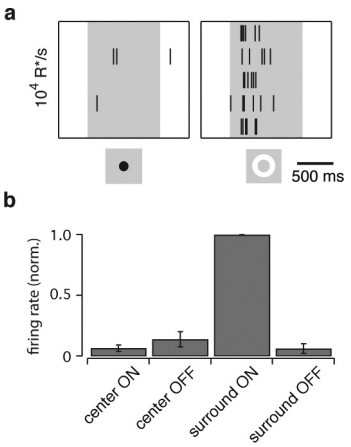
Extended Data Figure 3 | Responses of non-opponent J-RGCs in dorsal retina. a, Centre surround antagonism for non-opponent J-RGCs, at intermediate (top) and high (bottom) intensity. Each curve shows the peak firing rate in response to flashing spots of increasing size, measured separately at light onset (ON) and offset (OFF). Data were normalized for each cell and averaged over 7–14 cells (mean ± s.e.m.). **b**, **c**, Time course and spatial profile of the receptive field at different photopic intensities, averaged over 20 or 32 cells, respectively, and displayed as in Fig. 1e, f. **d**, Time course of excitatory (left) and inhibitory (right) conductance changes from stimulation of the S (blue trace) and M/rod-pigments (green) in the receptive field centre (top) and surround (bottom) (non-opponent cells $n = 5$, displayed as in Fig. 3a). **e**, Opsin space histogram for centre (top) and surround (bottom) currents (black: excitation, grey: inhibition). Note similarity to results from opponent J-RGCs (Fig. 3), except that the centre is driven by M pigment, as expected given the paucity of S cones in the dorsal region. The surround again has a pure M spectrum and produces both presynaptic and post-synaptic inhibition, with dynamics that are virtually identical to the signals in ventral J-RGCs.



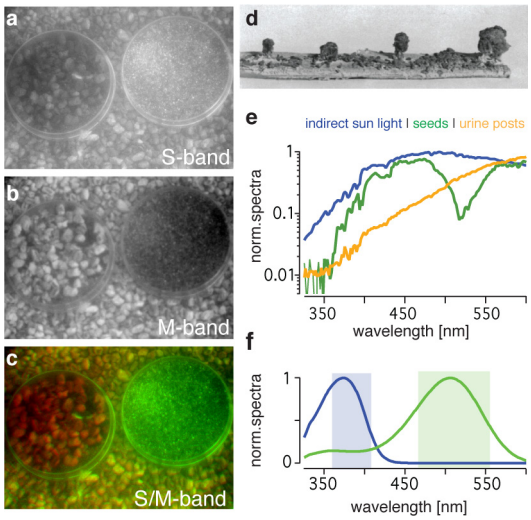
Extended Data Figure 4 | J-RGC current responses to flashed spots and annuli. a, Inhibitory and excitatory currents of a spectrally opponent J-RGC to a flashed spot (top, 250 μm diameter) and annulus (bottom, 2,000 μm and 350 μm for outer and inner diameter, respectively) centred on the receptive field using UV, green, or white (green + UV) light. **b**, **c**, Peak currents measured to a white flashed spot and annulus (**b**, inhibition, ON spot & OFF annulus; **c**, excitation, OFF spot & ON annulus; stimulus dimensions as in a) in control (black), picrotoxin (PTX, 100 μM; red), and combined (brown) PTX (100 μM) and strychnine (STR, 10 μM). Circles: individual cells; means ± s.e.m.; n.s.: not significant, $*P < 0.05$, $***P < 0.001$, one-way ANOVA. Note the synaptic currents in J-RGCs are systematically smaller (for example, excitatory current from centre stimulation = 10–20 pA) than those measured in other RGC types (for example, 500–1,000 pA typical in sustained alpha cells) during the same recording session.



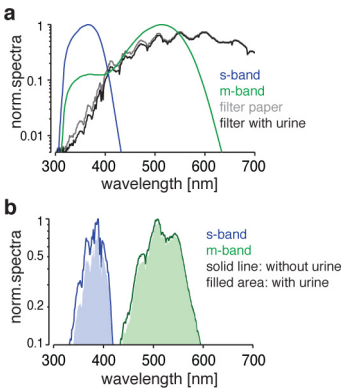
Extended Data Figure 5 | Synaptic pathways for spectral opponency (single cell examples). a, **d**, **g**, **j**, **m**, Inhibitory and excitatory currents during white-noise flicker stimulation of three different J-RGCs before and after drug application (shaded grey depicts the start of white-noise stimulus; **a**, **d**, L-AP4 11 μM; **g**, **j**, PTX 100 μM; **m**, L-AP4 11 μM, PTX 100 μM and STR 10 μM). **b**, **e**, **h**, **k**, **n**, Single cell excitatory and inhibitory current distribution under white-noise stimulation (before and during drug application). Inhibitory current distribution is dramatically narrowed during L-AP4 and PTX application, the excitatory current distribution remains comparatively unaltered. **c**, **f**, **i**, **l**, **o**, Single cell visual sensitivity of synaptic currents recorded from the respective J-RGC recordings in the left panels (as in Fig. 3a). Excitatory and inhibitory conductances are driven by stimulation of the centre (left) or surround (right) of the receptive field. Each curve represents the sensitivity of the conductance to stimulation of the M/rod pigment (green) or S pigment (blue) at various times in the past (see Methods).



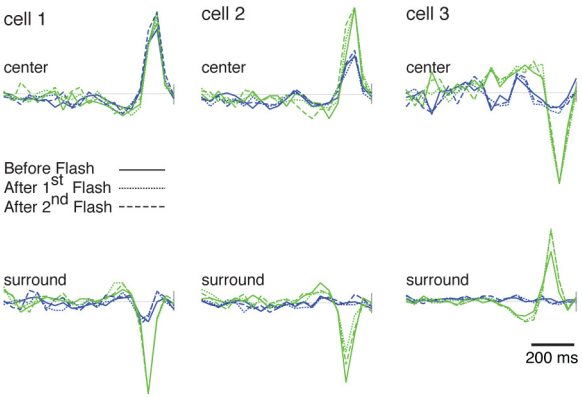
Extended Data Figure 6 | J-RGCs in a mutant retina with silenced cones. Responses of J-RGCs in the ventral retina of homozygous *Gnat2*^{cpfl3} mutant mice. Flashing spots and annuli as in Fig. 1d. **a**, Raster graph of spiking for one sample neuron. **b**, Summary of responses from 5 neurons. Firing rate normalized to that under 'surround ON' stimulation. Note little or no response to centre stimulation. Compare to wild-type retina in Figs 1d and 2d and Extended Data Fig. 1.



Extended Data Figure 7 | Spectrally opponent features in the environment. Dried mouse urine and plant seeds have high S-M chromatic contrast. **a**–**c**, On a background of clean mouse bedding are two dishes containing bedding soiled with urine (left) and a mix of plant seeds (right). Photographs used a band-pass filter in the ultraviolet (**a**) and in the green (**b**); **c** merges the two using red-green encoding. **d**, Close-up view of mouse urinating posts (reproduced with permission from ref. 26; 2.5–3.5 cm high). **e**, Normalized spectra of indirect sunlight and light reflected from a mixture of untreated plant seeds and urine posts. **f**, Pigment absorption curves for S opsin (blue) and M opsin (green) multiplied by the ocular transmission spectrum (see Methods). Shaded region indicates pass band of the filters used for **a**–**c**.



Extended Data Figure 8 | UV-green colour signature of urine. a, Pigment absorption curves for S opsin (blue) and M opsin (green) multiplied by the ocular transmission spectrum (see Methods) and spectra of light reflected from a Whatman filter paper, with or without dried urine marks, under indirect sunlight. **b**, Normalized curves of the product between the above absorption and reflectance spectra. Blue: S opsin, green: M opsin, solid line: clean filter paper, filled area: filter paper with urine. Note that the reduction in the S-band is 27.4%, compared to only 8.2% for the M-band.



Extended Data Figure 9 | Effects of the light flash exposure for fluorescent targeting. Temporal filter in the receptive field centre and surround for S opsin and M/rod opsin (as in Fig. 1f; blue and green traces, respectively) of three RGCs taken before and after one or two brief light flashes. These neurons were targeted blindly and are therefore not of the J-RGC type. They rely on both rods and cones (note different spectral sensitivity in the centre and surround). Yet their response properties were not altered by the brief flashes.

LETTER

doi:10.1038/nature17630

TAM receptors regulate multiple features of microglial physiology

Lawrence Fourgeaud[1]*, Paqui G. Través[1,2]*, Yusuf Tufail[3], Humberto Leal-Bailey[1,4], Erin D. Lew[1], Patrick G. Burrola[1], Perri Callaway[1], Anna Zagórska[1], Carla V. Rothlin[5], Axel Nimmerjahn[3] & Greg Lemke[1,6]

Microglia are damage sensors for the central nervous system (CNS), and the phagocytes responsible for routine non-inflammatory clearance of dead brain cells[1]. Here we show that the TAM receptor tyrosine kinases Mer and Axl[2] regulate these microglial functions. We find that adult mice deficient in microglial Mer and Axl exhibit a marked accumulation of apoptotic cells specifically in neurogenic regions of the CNS, and that microglial phagocytosis of the apoptotic cells generated during adult neurogenesis[3,4] is normally driven by both TAM receptor ligands Gas6 and protein S[5]. Using live two-photon imaging, we demonstrate that the microglial response to brain damage is also TAM-regulated, as TAM-deficient microglia display reduced process motility and delayed convergence to sites of injury. Finally, we show that microglial expression of Axl is prominently upregulated in the inflammatory environment that develops in a mouse model of Parkinson's disease[6]. Together, these results establish TAM receptors as both controllers of microglial physiology and potential targets for therapeutic intervention in CNS disease.

Microglia, the tissue macrophages of the brain and spinal cord, have fundamental roles in CNS homeostasis. They are mobilized in response to nearly any CNS perturbation, and can act to both resolve and exacerbate CNS disease[1,7]. Their importance notwithstanding, the signalling systems that regulate microglial function are only beginning to be deciphered. We asked whether the TAM receptor tyrosine kinases might comprise one such system. These receptors, Tyro3, Axl, and Mer, regulate the innate immune response in dendritic cells and macrophages[2,8,9], mediate the engulfment of apoptotic cells by phagocytes[10–12], promote the infection of cells by enveloped viruses[13], and contribute to the growth and metastasis of human cancers[14]. In the CNS, Tyro3 is abundant in neurons[15,16], whereas Mer and Axl are present in microglia[17–19].

Microglia express the fractalkine receptor Cx3cr1 and the ionized calcium-binding adaptor Iba1[20,21]. We therefore used Mer, Axl, glial fibrillary acidic protein (GFAP), and S100b antibodies to stain brain sections from *Cx3cr1*$^{GFP/+}$ and *S100b*$^{GFP/+}$ adult mice, which express GFP in microglia and astrocytes, respectively[22] (see Methods). Mer co-localized with GFP$^+$ microglia and not with GFAP$^+$ astrocytes or S100b$^+$ cells in *Cx3cr1*$^{GFP/+}$ mice (Extended Data Fig. 1a, b); and with Iba1 in wild-type and *S100b*$^{GFP/+}$ mice (Extended Data Fig. 1c). We do not exclude Mer expression in adult astrocytes (as seen for neonatal astrocytes[23]) at levels below immunohistochemical detection, but microglia abundantly express *Mertk* mRNA[17], and Mer plus CD64 is now the definitive marker pair for all tissue macrophages[17]. We detected only low Axl expression in the CNS. As shown below, this expression is elevated in inflammatory environments.

The best-studied role for TAM receptors is in the phagocytic clearance of apoptotic cells[2,10–12]. We therefore asked if Mer and Axl were required for clearance of the apoptotic cells generated during adult neurogenesis, which occurs in the subgranular zone of the dentate gyrus and the subventricular zone (SVZ) abutting the lateral ventricle, and produces neurons that integrate into the hippocampus and olfactory bulb, respectively[3]. We used the CLARITY imaging method with *Cx3cr1*$^{GFP/+}$ mice to show that the SVZ, like the rest of the CNS[1], is tiled by microglia (see Supplementary Video 1). During neurogenesis in the mouse subgranular zone, ~80% of cells die within 8 days of their birth[4]. These dead cells are rapidly cleared by microglia, such that apoptotic cells are difficult to detect in a healthy brain[4]. Indeed, when we examined SVZ sections from *Cx3cr1*$^{GFP/+}$ brains using cleaved caspase 3 (cCasp3) as a marker of apoptotic cells[5], we were unable to detect even a single cCasp3$^+$ cell in many sections (Fig. 1a). In marked contrast, the SVZ from *Axl*$^{-/-}$*Mertk*$^{-/-}$*Cx3cr1*$^{GFP/+}$ mice was studded with cCasp3$^+$ cells (Fig. 1a), which were negative for the neuronal marker NeuN (Fig. 1a). Uncleared apoptotic cells extended into the *Axl*$^{-/-}$*Mertk*$^{-/-}$ rostral migratory stream (RMS), the pathway through which newborn cells migrate to the olfactory bulb (Fig. 1b). Apoptotic cells were confined to *Axl*$^{-/-}$*Mertk*$^{-/-}$ neurogenic regions, however, and were not seen elsewhere in the CNS (Extended Data Fig. 2). Microglia in the *Axl*$^{-/-}$*Mertk*$^{-/-}$ SVZ and RMS displayed elevated expression of GFP (controlled by the *Cx3cr1* promoter), Iba1, and Siglec-1 (CD169), as well as an 'activated amoeboid' morphology[1] (Fig. 1a, c).

Consistent with the minimal expression of Axl, no accumulation of apoptotic cells was detected in the *Axl*$^{-/-}$ SVZ (Extended Data Fig. 3a). In contrast, the *Mertk*$^{-/-}$ SVZ contained many cCasp3$^+$ apoptotic cells, although this number was around fourfold lower than that seen in *Axl*$^{-/-}$*Mertk*$^{-/-}$ double mutants (Extended Data Fig. 3a). We counted 733 ± 359 (± s.e.m.) cCasp3$^+$ cells per mm^2 in sections of the *Mertk*$^{-/-}$ SVZ, and 2942 ± 262 cCasp3$^+$ cells per mm^2 in the *Axl*$^{-/-}$*Mertk*$^{-/-}$ SVZ. This synergistic effect of an *Axl* mutation on the background of an existing *Mertk* mutation has been noted previously[2,5,8,9,12].

We demonstrated that accumulation of apoptotic cells in the *Mertk*$^{-/-}$ SVZ and RMS is due to the loss of Mer specifically from microglia, by analysing a new mouse line carrying conditional floxed alleles of the *Mertk* gene (Extended Data Fig. 4; see Methods) crossed to a tamoxifen-inducible oestrogen receptor (ER) Cre driver controlled by the *Cx3cr1* promoter[24]. In the absence of tamoxifen (upon vehicle injection alone), Mer was present in *Cx3cr1*$^{CreER/+}$*Mertk*$^{fl/fl}$ Iba1$^+$ microglia (Extended Data Fig. 5a), and there were no cCasp3$^+$ apoptotic cells in the SVZ or RMS (Fig. 1d). However, 1 week after tamoxifen injection, microglial Mer expression was lost (Extended Data Fig. 5a), and accumulation of apoptotic cells, comparable to that seen in *Mertk*$^{-/-}$ mice, was detected in the *Cx3cr1*$^{CreER/+}$*Mertk*$^{fl/fl}$ SVZ and RMS (Fig. 1d). Microglia remained

[1]Molecular Neurobiology Laboratory, The Salk Institute for Biological Studies, La Jolla, California 92037, USA. [2]Instituto de Investigaciones Biomédicas Alberto Sols (CSIC-UAM), Madrid 28029, Spain. [3]Waitt Advanced Biophotonics Center, The Salk Institute for Biological Studies, La Jolla, California 92037, USA. [4]Joint Master in Neuroscience Program, University of Strasbourg, Strasbourg 67081, France. [5]Department of Immunobiology, Yale University School of Medicine, New Haven, Connecticut 06520, USA. [6]Immunobiology and Microbial Pathogenesis Laboratory, The Salk Institute for Biological Studies, La Jolla, California 92037, USA.
*These authors contributed equally to this work.

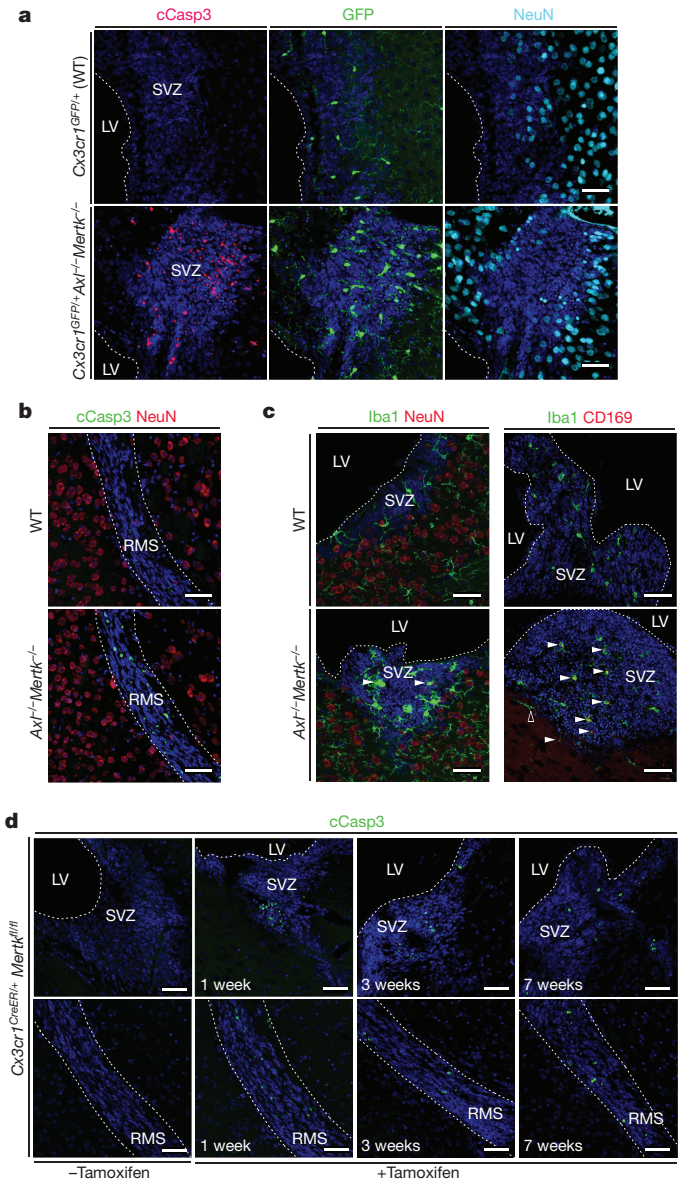


Figure 1 | TAM signalling mediates microglial phagocytosis of apoptotic cells in brain neurogenic regions. a, SVZ sections adjacent to the lateral ventricle (LV) of wild-type (WT) or $Axl^{-/-}Mertk^{-/-}Cx3cr1^{GFP/+}$ brains visualized for GFP (green), cCasp3 (magenta), and NeuN (cyan). **b**, RMS of wild-type and $Axl^{-/-}Mertk^{-/-}$ brains immunostained for cCasp3 (green) and NeuN (red). **c**, Immunostaining of wild-type and $Axl^{-/-}Mertk^{-/-}$ SVZ with anti-Iba1 (green) and anti-NeuN (red), or anti-Iba1 (green) and anti-CD169 (red). Arrowheads mark $Iba1^+$ microglia with an amoeboid morphology (lower left) and $Iba1^+CD169^+$ double-positive microglia (lower right); open arrowhead is an $Iba1^+CD169^-$ cell outside the SVZ. **d**, No $cCasp3^+$ apoptotic cells accumulate in the SVZ or RMS of $Cx3cr1^{CreER/+}Mertk^{fl/fl}$ mice 1 week after vehicle injection (−tamoxifen), but many are evident in the SVZ and RMS at 1, 3, and 7 weeks after injection with vehicle and tamoxifen to induce Cre expression in $Cx3cr1^+$ microglia (+tamoxifen). All sections in **a**–**d** are co-stained with nuclear Hoechst 33258 (blue). Representative images from analyses performed in 3 (**a**, **b**, **d**) and 2 (**c**) mice. Scale bars, 50 μm.

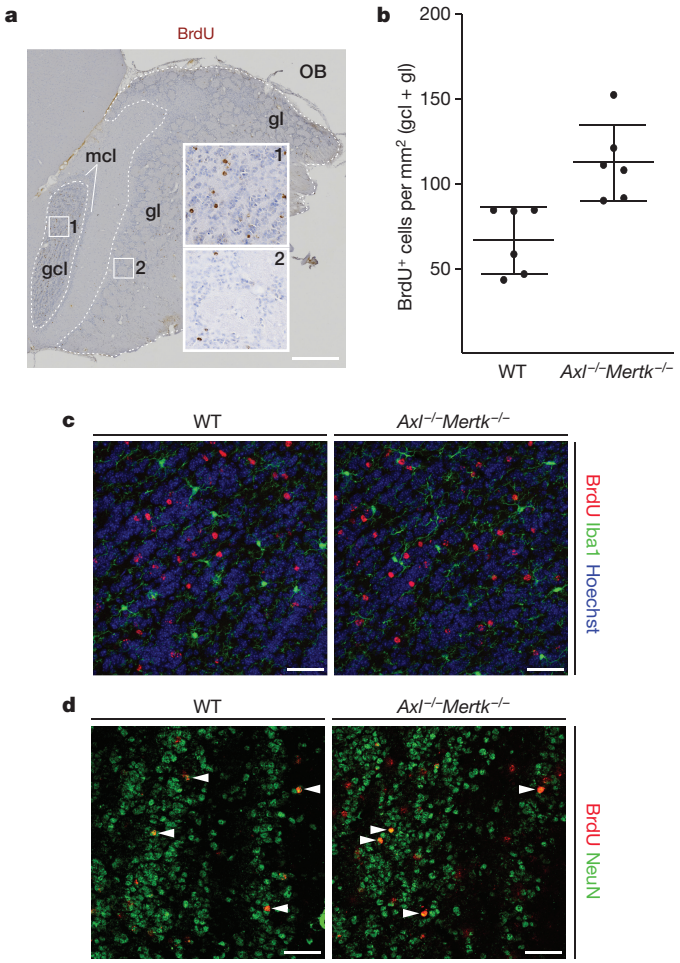


Figure 2 | TAM signalling may mediate death by phagocytosis. a, $Axl^{-/-}Mertk^{-/-}$ olfactory bulb section five weeks after BrdU pulse labelling, visualized with an anti-BrdU antibody (brown). The granule cell layer (gcl), glomerular layer (gl), and mitral cell layer (mcl) are indicated, and regions of the granule cell layer (1) and glomerular layer (2) are enlarged. Scale bar, 500 μm. **b**, Quantification of $BrdU^+$ cells per mm^2 in the granule cell layer and glomerular layer of 6 wild-type versus 6 $Axl^{-/-}Mertk^{-/-}$ mice. Graph plots average ± s.e.m.; two-tailed unpaired Mann–Whitney $P = 0.002$. **c**, $BrdU^+$ cells in the granule cell layer 35 days after injection of BrdU (red) are negative for Iba1 (green) in wild-type and $Axl^{-/-}Mertk^{-/-}$ mice. **d**, Similar gcl sections stained with anti-BrdU (red) and NeuN (green). Arrowheads mark $NeuN^+BrdU^+$ cells. Sections in **c** were co-stained with Hoechst 33258. Scale bars (**c**, **d**), 50 μm. Representative images of $n = 2$ per genotype.

Mer-negative (Extended Data Fig. 5a) and accumulation of apoptotic cells was maintained at 3 and 7 weeks after tamoxifen injection (Fig. 1d), by which time most $Cx3cr1^{CreER/+}$ gene-deleted cells outside the CNS have been replaced by monocytes and/or haematopoietic progenitors[24]. Mer expression in brain microvascular endothelial cells[23], an important Mer reservoir in the CNS[25], persisted following tamoxifen treatment (Extended Data Fig. 5b).

We assessed the consequences of defective clearance of apoptotic cells on neurogenesis by pulse labelling dividing cells in the SVZ of adult mice with bromodeoxyuridine (BrdU), and then counting $BrdU^+$ cells that had migrated to the granule cell and glomerular layers of the olfactory bulb 35 days after the pulse (Fig. 2a, b). We found that accumulation of apoptotic cells in the $Axl^{-/-}Mertk^{-/-}$ SVZ did not reduce the number of $BrdU^+$ cells in the olfactory bulb, none of which were apoptotic (Extended Data Fig. 2c) or microglia (Fig. 2c). Indeed, we observed a notable ~70% increase in the number of $BrdU^+$ cells in the $Axl^{-/-}Mertk^{-/-}$ olfactory bulb relative to the wild type (Fig. 2b). This translated to an increased cellular density in the combined granule cell and glomerular layers of the $Axl^{-/-}Mertk^{-/-}$ olfactory bulb, from 87.7 ± 2.2 nuclei per $10^4\,\mu m^2$ (± s.e.m.) in wild type to 99.8 ± 2.4 for the double mutants ($n = 6$ for both genotypes; $P = 0.004$). These results are consistent with the possibility that a fraction of phosphatidylserine (PtdSer)-expressing, but nonetheless viable, SVZ-derived cells are normally 'eaten alive' by microglia, in a process termed 'phagoptosis'[26], that this process occurs continuously in a non-pathogenic environment,

and that it is TAM-dependent. Many of the BrdU$^+$ cells that had migrated to the *Axl*$^{-/-}$*Mertk*$^{-/-}$ olfactory bulb were NeuN$^+$ (Fig. 2d), and some expressed markers appropriate to their location (Extended Data Fig. 6a, b).

Genetic analyses *in vivo* indicated that both protein S (Pros1) and Gas6 function as Mer agonists for engulfment of apoptotic cells by microglia. The *Gas6*$^{-/-}$ SVZ displayed a wild-type phenotype (Extended Data Fig. 3b), as did the SVZ of *Gas6*$^{-/-}$*Pros1*$^{fl/-}$ mice, in which one *Pros1* allele is floxed with loxP sites and the other is inactivated[27]. (The complete *Pros1*$^{-/-}$ knockout is embryonic lethal[27].) In contrast, *Gas6*$^{-/-}$*Mertk*$^{-/-}$ mice displayed a marked accumulation of cCasp3$^+$ cells in the SVZ and RMS, comparable to that seen in *Axl*$^{-/-}$*Mertk*$^{-/-}$ mice (Extended Data Fig. 3c). We counted 3,638 ± 282 (± s.e.m.) cCasp3$^+$ cells per mm^2 in the *Gas6*$^{-/-}$*Mertk*$^{-/-}$ mice SVZ. This is consistent with the fact that Pros1, the only TAM ligand remaining in the *Gas6*$^{-/-}$*Mertk*$^{-/-}$ mice, does not activate Axl[5], the only microglial TAM receptor remaining in these mice. Thus, as for Mer-dependent phagocytosis in the retina[10], only half of the wild-type level of only a single TAM ligand was sufficient to drive wild-type levels of microglial phagocytosis.

We also quantified phagocytosis of apoptotic cells by microglia cultured from *Cx3cr1*$^{GFP/+}$ mice[12] (Extended Data Fig. 7a). When incubated with apoptotic cells[12] in medium containing 10% serum, where Pros1 is present at ~30 nM, wild-type microglia were exceptionally active phagocytes (Extended Data Fig. 7b). Phagocytosis was reduced in *Axl*$^{-/-}$*Mertk*$^{-/-}$ cells (Extended Data Fig. 7b). In serum-free medium, phagocytosis of apoptotic cells was further reduced, which was mostly TAM-dependent (Extended Data Fig. 7c, d). (This TAM dependence is consistent with the fact that microglia express endogenous *Gas6* and *Pros1* mRNA[17].) When we supplemented serum-free medium with Pros1 or Gas6, we found that both ligands stimulated TAM-dependent phagocytosis (Extended Data Fig. 7c, d). Cultured astrocytes engulfed less apoptotic cells than microglia, and this phagocytosis could be stimulated only modestly by Gas6 (Extended Data Fig. 7e). Stimulation of microglial phagocytosis by both Gas6 and Pros1 demonstrates that it is mediated principally by Mer, as Axl is activated only by Gas6[5,12].

Although microglia in an uninjured brain are essentially fixed in position, their processes are in constant motion, and survey the entirety of the CNS parenchyma every few hours[28]. As process extension is also required for phagocytosis[12], we asked if TAM signalling regulates microglial extension velocity. We used *in vivo* two-photon microscopy to measure the movement of microglial processes outside of the neurogenic regions, in the visual cortex of wild-type *Cx3cr1*$^{GFP/+}$ and *Axl*$^{-/-}$*Mertk*$^{-/-}$*Cx3cr1*$^{GFP/+}$ mice (Supplementary Videos 2 and 3). Selected video stills, with individual wild-type and *Axl*$^{-/-}$*Mertk*$^{-/-}$ processes tracked during imaging, are shown in Fig. 3a. These measurements demonstrated that *Axl*$^{-/-}$*Mertk*$^{-/-}$ microglia, in an uninjured brain, display a ~19% reduction in process extension velocity relative to wild type (Fig. 3b). We also assessed whether TAM signalling was required for microglial responses to injury. We disrupted the blood-brain barrier at the level of individual capillaries with a laser lesion[28], and then measured the velocity of microglial process extension towards the lesion site using live two-photon imaging (Supplementary Videos 4 and 5). Selected video stills, with individual extensions towards the lesion tracked in wild-type and *Axl*$^{-/-}$*Mertk*$^{-/-}$*Cx3cr1*$^{GFP/+}$ brains, are shown in Fig. 3c. We found that extension towards the laser lesion was ~39% slower in the *Axl*$^{-/-}$*Mertk*$^{-/-}$ microglia (Fig. 3d). These results demonstrate that routine microglial process activity and the response to injury are both regulated by TAM. They are consistent with the finding that *Mertk*$^{-/-}$ macrophages exhibit compromised cellular migration and a disrupted cytoskeleton *in vitro*[29].

For many macrophages, Axl and Mer segregate to inflammatory and tolerogenic environments, respectively[12]. Inflammatory stimuli such as polyinosinic–polycytidylic acid (poly(I:C)) and interferon γ(IFNγ) upregulate Axl expression, whereas immunosuppressive drugs such as

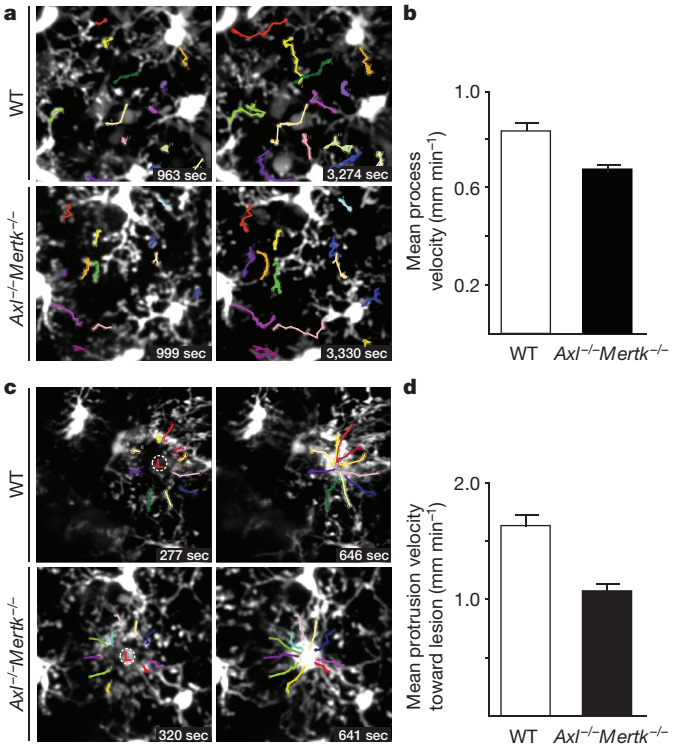


Figure 3 | TAM signalling regulates microglial process extension velocity and response to vascular injury. a, Two video stills for wild-type and *Axl*$^{-/-}$*Mertk*$^{-/-}$ *Cx3cr1*$^{GFP/+}$ mice, with tracking of individual GFP-labelled processes (colour-coded) in the unperturbed visual cortex, by live two-photon imaging. Stills are from Supplementary Videos 2 (wild type) and 3 (*Axl*$^{-/-}$*Mertk*$^{-/-}$), and indicated times are from the start of the video. **b**, Mean process velocity (± s.e.m.) in the absence of perturbation. $n = 53$ measurements in 3 wild-type mice, and $n = 42$ in 3 *Axl*$^{-/-}$*Mertk*$^{-/-}$ mice; two-tailed unpaired Mann–Whitney $P = 0.0004$. **c**, Two video stills for wild-type and *Axl*$^{-/-}$*Mertk*$^{-/-}$ *Cx3cr1*$^{GFP/+}$ mice, illustrating microglial process tracking (colour-coded) towards a laser-induced rupture of a brain microvessel (circled L) by two-photon imaging. Stills are from Supplementary Videos 4 (wild-type) and 5 (*Axl*$^{-/-}$*Mertk*$^{-/-}$), and indicated times are from the generation of the laser lesion. **d**, Mean process extension velocity (± s.e.m.) towards the lesion site. $n = 20$ measurements in 2 wild-type mice and $n = 30$ in 3 *Axl*$^{-/-}$*Mertk*$^{-/-}$ mice; two-tailed unpaired Mann–Whitney $P = < 0.0001$.

dexamethasone upregulate Mer[12]. We saw similar responses in cultured microglia (Extended Data Fig. 8a, b). As Axl is an inflammatory marker, we asked whether its microglial expression might be elevated in neurodegenerative disease[1], and examined a transgenic mouse model of Parkinson's disease. In this model, an alanine 53 to threonine (A53T) mutated form of human α-synuclein (*SNCA*^{A53T}), which leads to a hereditary form of Parkinson's disease, is expressed in neurons, most prominently in the spinal cord, under the control of the mouse *Thy1* promoter[6]. This transgenic expression results in late-onset neurodegeneration, and death at 8–10 months[6]. Measurement of a panel of inflammatory marker mRNAs demonstrated that the spinal cords, and to a lesser extent the brains, of aged *Thy1-SNCA*^{A53T} transgenic mice displayed elevation of these markers, whereas inflammation was undetectable in the spleen (Fig. 4a, Extended data Fig. 8c).

The aged transgenic spinal cord, where *SNCA*^{A53T} expression is high (ref. 29 and Fig. 4b), showed markedly elevated expression of Iba1 (Fig. 4b). We also detected upregulation of both Axl and soluble Axl (sAxl) ectodomain, an inflammatory marker[12], in the transgenic cord (Fig. 4c and Extended Data Fig. 8d). In contrast, Axl upregulation was undetectable in *Thy1-SNCA*^{A53T} spleen and minimal in brain (Extended Data Fig. 8d). No change in Mer expression was detected in the spleen

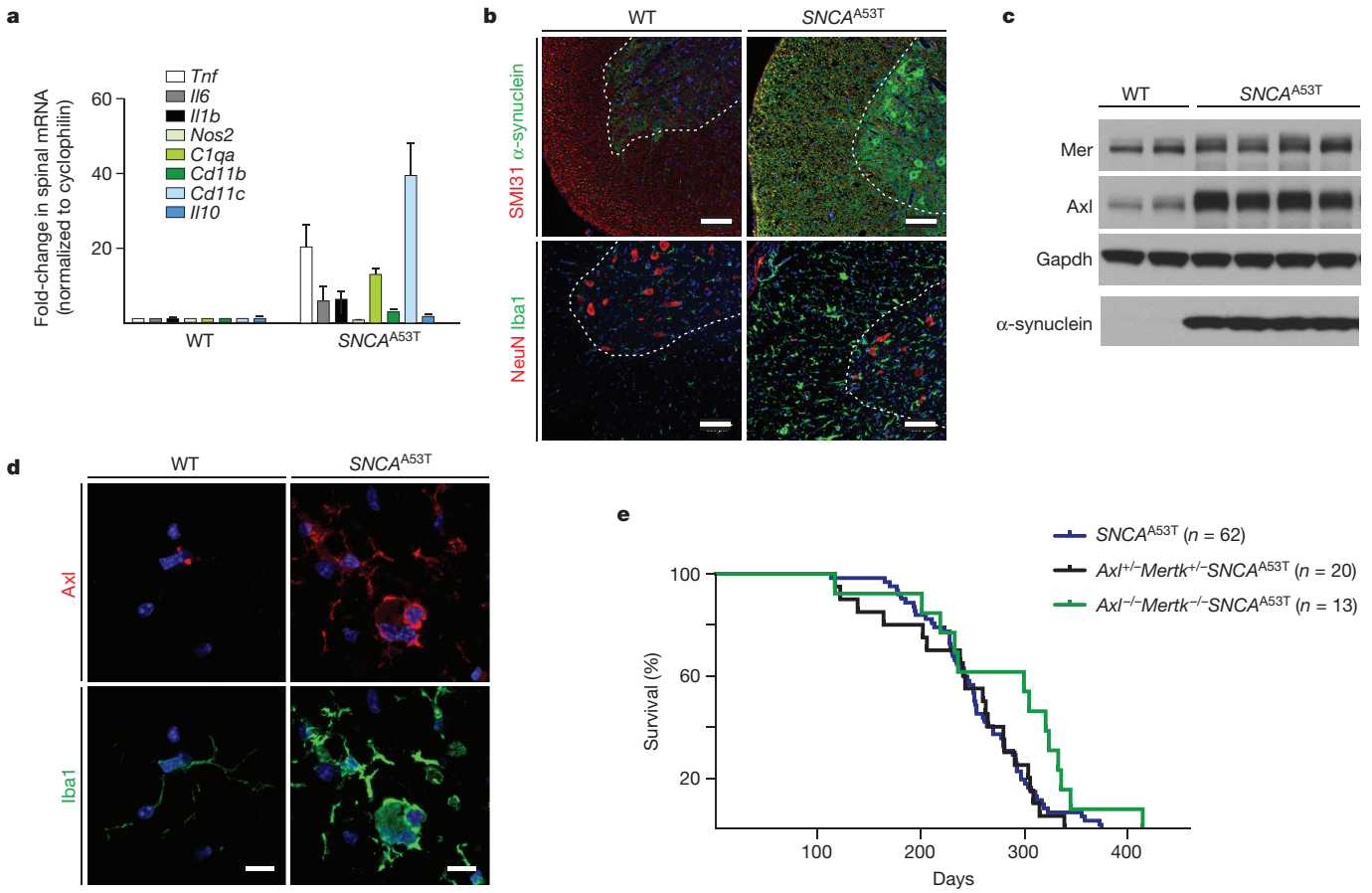


Figure 4 | Microglial Axl is upregulated in a mouse model of Parkinson's disease. a, Comparison of the mean expression (± s.e.m.) of the indicated inflammatory mediator/marker mRNAs in the spinal cords of 3 wild-type and 3 *Thy1-SNCA*^{A53T} (*SNCA*^{A53T}) mice at 8–10 months of age. **b**, Sections from the spinal cords of aged (8–9 month) wild-type and *SNCA*^{A53T} mice, immunostained for phosphorylated neurofilament (SMI31) and α-synuclein, or NeuN and Iba1. The α-synuclein antibody recognizes both the endogenous mouse protein and the transgenic human protein. Scale bars, 100 μm. **c**, Western blots of spinal cord extracts from 2 wild-type mice (lanes 1, 2) and 4 *SNCA*^{A53T} mice (lanes 3–6) for the indicated proteins at 9–10 months of age, with Gapdh as a loading control. **d**, Sections from wild-type and *SNCA*^{A53T} spinal cords immunostained with anti-Axl and anti-Iba1 antibodies. Scale bar, 10 μm. **e**, Kaplan–Meier survival curves for mice of the indicated genotypes. $n = 62$ *SNCA*^{A53T} mice; 20 *Axl*$^{+/-}$*Mertk*$^{+/-}$*SNCA*^{A53T}; and 13 *Axl*$^{-/-}$*Mertk*$^{-/-}$*SNCA*^{A53T} mice. log-rank (Mantel–Cox) test $P = 0.72$ between *SNCA*^{A53T} and *Axl*$^{+/-}$*Mertk*$^{+/-}$*SNCA*^{A53T}; $P = 0.04$ between *SNCA*^{A53T} and *Axl*$^{-/-}$*Mertk*$^{-/-}$*SNCA*^{A53T}. Representative images from $n = 2$ wild-type and 3 *SNCA*^{A53T} mice (**b** and **d**).

of the transgenic mice (Extended Data Fig. 8d), with only a very modest increase in the spinal cord (Fig. 4c and Extended Data Fig. 8e). Axl induction in the *Thy1-SNCA*^{A53T} spinal cord was exclusively associated with Iba1^{+} microglia (Fig. 4d). Expression of *SNCA*^{A53T} in spinal motor neurons (ref. 29 and Fig. 4b) leads to progressive ataxia, paralysis, and death, with an onset at ~120 days in the transgenic population (Fig. 4e). A 50% reduction in Mer and Axl resulted in no change in this time course, but the loss of both receptors modestly extended survival (Fig. 4e). We do not know the reason for this modest life extension; however, we speculate that wild-type microglia may execute TAM-dependent 'phagoptotic' engulfment[26] of distressed, PtdSer-displaying motor neurons, thereby speeding up the death of the mice.

Together, the above results identify Mer and Axl as regulators of multiple features of microglial physiology. The elevation in microglial Axl that we document in the *Thy1-SNCA*^{A53T} spinal cord is in keeping with the demonstration that Axl is an inflammatory response receptor in macrophages[12], and that elevated levels of sAxl are observed during multiple human disease and trauma states (ref. 8, and references therein). In this regard, we note that a recent longitudinal study in humans has identified elevated Axl in CSF as among the most reliable indicators of the early appearance of Aβ pathology and the subsequent development of Alzheimer's disease[30].

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 26 May 2015; accepted 1 March 2016.
Published online 6 April 2016.

1. Ransohoff, R. M. & Cardona, A. E. The myeloid cells of the central nervous system parenchyma. *Nature* **468,** 253–262 (2010).
2. Lemke, G. Biology of the TAM receptors. *Cold Spring Harb. Perspect. Biol.* **5,** a009076 (2013).
3. Aimone, J. B. *et al.* Regulation and function of adult neurogenesis: from genes to cognition. *Physiol. Rev.* **94,** 991–1026 (2014).
4. Sierra, A. *et al.* Microglia shape adult hippocampal neurogenesis through apoptosis-coupled phagocytosis. *Cell Stem Cell* **7,** 483–495 (2010).
5. Lew, E. D. *et al.* Differential TAM receptor-ligand-phospholipid interactions delimit differential TAM bioactivities. *eLife* **3,** e03385 (2014).
6. Chandra, S., Gallardo, G., Fernandez-Chacon, R., Schluter, O. M. & Sudhof, T. C. α-Synuclein cooperates with CSPα in preventing neurodegeneration. *Cell* **123,** 383–396 (2005).
7. Ginhoux, F. *et al.* Fate mapping analysis reveals that adult microglia derive from primitive macrophages. *Science* **330,** 841–845 (2010).
8. Lu, Q. & Lemke, G. Homeostatic regulation of the immune system by receptor tyrosine kinases of the Tyro 3 family. *Science* **293,** 306–311 (2001).
9. Rothlin, C. V., Ghosh, S., Zuniga, E. I., Oldstone, M. B. & Lemke, G. TAM receptors are pleiotropic inhibitors of the innate immune response. *Cell* **131,** 1124–1136 (2007).
10. Burstyn-Cohen, T. *et al.* Genetic dissection of TAM receptor-ligand interaction in retinal pigment epithelial cell phagocytosis. *Neuron* **76,** 1123–1132 (2012).

11. Scott, R. S. *et al.* Phagocytosis and clearance of apoptotic cells is mediated by MER. *Nature* **411,** 207–211 (2001).
12. Zagórska, A., Través, P. G., Lew, E. D., Dransfield, I. & Lemke, G. Diversification of TAM receptor tyrosine kinase function. *Nature Immunol.* **15,** 920–928 (2014).
13. Bhattacharyya, S. *et al.* Enveloped viruses disable innate immune responses in dendritic cells by direct activation of TAM receptors. *Cell Host Microbe* **14,** 136–147 (2013).
14. Zhang, Z. *et al.* Activation of the AXL kinase causes resistance to EGFR-targeted therapy in lung cancer. *Nature Genet.* **44,** 852–860 (2012).
15. Lai, C. & Lemke, G. An extended family of protein-tyrosine kinase genes differentially expressed in the vertebrate nervous system. *Neuron* **6,** 691–704 (1991).
16. Prieto, A. L., O'Dell, S., Varnum, B. & Lai, C. Localization and signaling of the receptor protein tyrosine kinase Tyro3 in cortical and hippocampal neurons. *Neuroscience* **150,** 319–334 (2007).
17. Gautier, E. L. *et al.* Gene-expression profiles and transcriptional regulatory pathways that underlie the identity and diversity of mouse tissue macrophages. *Nature Immunol.* **13,** 1118–1128 (2012).
18. Grommes, C. *et al.* Regulation of microglial phagocytosis and inflammatory gene expression by Gas6 acting on the Axl/Mer family of tyrosine kinases. *J. Neuroimmune Pharmacol.* **3,** 130–140 (2008).
19. Ji, R. *et al.* TAM receptors affect adult brain neurogenesis by negative regulation of microglial cell activation. *J. Immunol.* **191,** 6165–6177 (2013).
20. Cardona, A. E. *et al.* Control of microglial neurotoxicity by the fractalkine receptor. *Nature Neurosci.* **9,** 917–924 (2006).
21. Ito, D. *et al.* Microglia-specific localisation of a novel calcium binding protein, Iba1. *Brain Res. Mol. Brain Res.* **57,** 1–9 (1998).
22. Jung, S. *et al.* Analysis of fractalkine receptor CX_3CR1 function by targeted deletion and green fluorescent protein reporter gene insertion. *Mol. Cell. Biol.* **20,** 4106–4114 (2000).
23. Chung, W. S. *et al.* Astrocytes mediate synapse elimination through MEGF10 and MERTK pathways. *Nature* **504,** 394–400 (2013).
24. Parkhurst, C. N. *et al.* Microglia promote learning-dependent synapse formation through brain-derived neurotrophic factor. *Cell* **155,** 1596–1609 (2013).
25. Miner, J. J. *et al.* The TAM receptor Mertk protects against neuroinvasive viral infection by maintaining blood-brain barrier integrity. *Nature Med.* **21,** 1464–1472 (2015).
26. Brown, G. C. & Neher, J. J. Microglial phagocytosis of live neurons. *Nature Rev. Neurosci.* **15,** 209–216 (2014).
27. Burstyn-Cohen, T., Heeb, M. J. & Lemke, G. Lack of protein S in mice causes embryonic lethal coagulopathy and vascular dysgenesis. *J. Clin. Invest.* **119,** 2942–2953 (2009).
28. Nimmerjahn, A., Kirchhoff, F. & Helmchen, F. Resting microglial cells are highly dynamic surveillants of brain parenchyma *in vivo*. *Science* **308,** 1314–1318 (2005).
29. Tang, Y. *et al.* Mertk deficiency affects macrophage directional migration via disruption of cytoskeletal organization. *PLoS ONE* **10,** e0117787 (2015).
30. Mattsson, N. *et al.* CSF protein biomarkers predicting longitudinal reduction of CSF β-amyloid42 in cognitively healthy elders. *Transl. Psychiatry* **3,** e293 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by grants from the US National Institutes of Health (R01 NS085296 and R01 AI101400 to G.L., DP2 NS083038 and R01 NS085938 to A.N., R01 AI089824 to C.V.R., and P30CA014195 to the Salk Institute), the Leona M. and Harry B. Helmsley Charitable Trust (#2012-PG-MED002 to the Salk Institute), the Nomis, H. N. and Frances C. Berger, Fritz B. Burns, and HKT Foundations (to G.L.), and the Waitt, Rita Allen and Hearst Foundations (to A.N.); and by postdoctoral fellowships from the Marie Curie International Outgoing Fellowship Program (to P.G.T.), the Nomis Foundation (to A.Z. and E.D.L.), and the Howard Hughes Medical Institute Life Sciences Research Foundation (to Y.T.). We thank J. Hash for excellent technical assistance and J. Flynn for help with the CLARITY method.

Author Contributions L.F. and P.G.T. designed experiments, performed apoptotic cell, BrdU, immunohistochemical, and genetic analyses, and contributed equally to the paper; Y.T., L.F. and H.L.-B. performed and analysed *in vivo* two photon imaging; E.D.L. prepared TAM ligands; P.G.B. performed brain histology; P.C. analysed cytokine profiles; A.Z. analysed Axl expression in Parkinson's disease transgenics; C.V.R. provided floxed *Mertk* alleles; A.N. designed and implemented two-photon imaging; and G.L. designed experiments and wrote the paper. All authors edited the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to G.L. (lemke@salk.edu) or C.V.R. (carla.rothlin@yale.edu; for requests for floxed *Mertk* alleles).

METHODS

Mice. The *Axl*$^{-/-}$[31], *Mertk*$^{-/-}$[31], *Axl*$^{-/-}$*Mertk*$^{-/-}$[31], *Gas6*$^{-/-}$[31] *Pros1*$^{fl/fl}$*NesCre*[10,32], *Pros1*$^{fl/-}$*NesCre*[10], *Cx3cr1*$^{GFP/+}$[22], *Cx3cr1*^{CreER}[24], *S100b*$^{GFP/+}$[33] and *SNCA*^{A53T} (ref. 6) strains have been described previously. The *Mertk*$^{fl/fl}$ mouse line diagrammed in Extended Data Fig. 4 was generated by inGenious Targeting Laboratory (iTL, Ronkonkoma NY), using iTL C57Bl/6 embryonic stem (ES) cells. This line targets exon 18, a 137 nucleotide (nt) sequence that encodes residues W779–L824 within the Mer kinase domain. Cre-mediated deletion of this exon introduces a frame shift and a stop codon one amino acid downstream of exon 17. This truncated, kinase-dead protein and/or its mRNA are apparently unstable, as antibodies directed against the Mer extracellular domain do not detect a truncated protein upon Cre-mediated excision (see text). Deletion of exon 18 therefore effectively generates a protein null. The complete *Mertk* mouse knockout[31] deletes exon 17, a 160 nt sequence that encodes M725–V778 within the Mer kinase domain. (Exon 17 was numbered as exon 18 in the original description of the *Mertk* knockout allele[31].) This single exon deletion also introduces a frame shift (five amino acids downstream of exon 16), produces an unstable protein, and also results in a Mer protein null[12]. The Neo cassette was removed via Flp-mediated recombination by crossing high-percentage chimaeric mice to C57Bl/6 FLP mice. Neo deletion was confirmed by PCR. These *Mertk*$^{fl/fl}$ mice, together with PCR-based protocols for their genotyping, are available upon request from the Rothlin laboratory (contact C.V.R.). Recombination (inactivation) of the *Mertk*$^{fl/fl}$ allele in *Cx3cr1*$^{CreER/+}$*Mertk*$^{fl/fl}$ mice was achieved using tamoxifen injection. *Cx3cr1*$^{CreER/+}$*Mertk*$^{fl/fl}$ mice (16 weeks) received a dose (150 mg kg^{-1} body weight) of tamoxifen (Sigma) as a solution in corn oil (Sigma) by intraperitoneal (i.p.) injection. Control mice received an i.p. injection of vehicle (corn oil) alone. Mice were analysed for Mer expression and apoptotic-cell (cCasp3^{+} cell) accumulation 1 week, 3 weeks or 7 weeks after injection. Mice analysed at 1 week received a single dose of tamoxifen or oil; mice analysed at 3 and 7 weeks received two successive injections 48 h apart. All lines, with the exception of the *Mertk*$^{fl/fl}$ alleles, have been backcrossed for >9 generations to a C57BL/6 background. All animal procedures were conducted according to protocols approved by the Salk Institute Animal Care and Use Committee (Protocol No. 11-00051). Mice (both males and females) were randomly allocated to experimental groups (three to six mice per group) and investigators were blinded to group allocation during the experiment. Investigators were not blinded to sample identity. Group size was based on previous literature. No statistical methods were used to predetermine sample size.

Reagents and antibodies. Dexamethasone, 5-Bromo-2-deoxyuridine and DMSO were from Sigma-Aldrich. Poly(I:C) was from Invivogen. Lipopolysaccharide (LPS) (*Escherichia coli* serotype O55:B5) was from Enzo. IFN-γ was from BioVision. Purified human protein S was from Haematologic Technologies. Recombinant mouse Gas6 was produced as described previously[5]. Antibodies used were as follows: anti-Mer (AF591), anti-Axl (AF854), and anti-Gas6 (AF986) were from R&D Systems, anti-Mer (DS5MMER) from eBioscience, anti-Iba1 (019-19741) was from Wako, anti-GFAP (z0334) was from Dako, anti-Neurofilament H (SMI-31 NE1022), anti-NeuN (MAB377 A60), anti-Calretinin (AB1550), anti-Tyrosine Hydroxylase (MAB318; LNC1) and anti-GAPDH (MAB374; 6C5) were from Millipore, anti-α-synuclein (C-20-R sc-7011-R) and anti-Axl (M-20 sc-1097) were from Santa Cruz, anti-cCaspase 3 (Asp175) was from Cell Signaling, anti-ACSA-2 (clone IH3-18A3) was from Miltenyi Biotec, anti-CD169 (Siglec1; 3D6) and anti-BrdU (BU1/75 (ICR1) were from AbD serotec, and anti CD31 (ab28364) and anti-S100b (EP1576Y) were from Abcam. Secondary antibodies used for immunoblot analysis were horseradish-peroxidase-conjugated anti-goat (705-035-003) from Jackson ImmunoResearch, and anti-mouse (NA931V) and anti-rabbit (NA934V) from GE Healthcare. Secondary antibodies for immunocyto- and immunohistochemistry were fluorophore-conjugated anti-goat (A-11055 from Life Technologies, or 705-166-147 from Jackson ImmunoResearch), anti-rabbit (A-10040 or A-21206 from Life Technologies), and anti-mouse (A-11029 from Life Technologies, or 715-166-150 from Jackson ImmunoResearch).

Immunohistochemistry. Adult mice (3–6 month) were anaesthetized with 2.5% avertin in saline, perfused with 20 U ml^{-1} heparin in PBS, and subsequently with 4% PFA in PBS. Brain and spinal cords were collected, immersion-fixed overnight at 4 °C, infiltrated with 30% sucrose in PBS overnight at 4 °C, and flash-frozen in tissue freezing medium. Sections of 17 μm were cut, air-dried overnight at room temperature and subsequently processed for staining. Non-specific binding was blocked by 1 h incubation in blocking buffer (PBS containing 0.1% Tween-20, 5% donkey serum and 2% IgG-free BSA). Sections were incubated overnight at 4 °C with primary antibody (identified above) diluted in blocking buffer, then washed in PBS 0.1% Tween-20, and incubated for 2 h at 22–24 °C in the dark with Hoechst and fluorophore-coupled secondary antibodies diluted in blocking buffer. Sections were washed, sealed with Fluoromount-G (SouthernBiotech) and stored at 4 °C. Images were acquired with a Zeiss LSM 710 confocal microscope using Plan-Apochromat 40× and 63× objectives.

Quantification of apoptotic cells. Cleaved Casp3^{+} apoptotic cells were counted in four successive 17 μm sections that spanned the SVZ in three different mice for both the *Mertk*$^{-/-}$ and A*xl*$^{-/-}$*Mertk*$^{-/-}$ genotypes, and in two different mice for the *Mertk*$^{-/-}$*Gas6*$^{-/-}$ genotype. No cCasp3^{+} cells in excess of wild-type were observed in SVZ sections of any of the other genotypes analysed. The cross-sectional area of the SVZ was defined as the region of intense Hoechst 33258 staining, as illustrated in Figs 1a, c, d, and measured using ImageJ. Accumulation of apoptotic cells between the *Axl*$^{-/-}$*Mertk*$^{-/-}$ and *Mertk*$^{-/-}$*Gas6*$^{-/-}$ genotypes is not statistically different. Note that cCasp3 marks a subset of apoptotic cells.

BrdU pulse labelling. Three successive injections (50 mg kg^{-1} body weight) of 5-bromo-2-deoxyuridine (BrdU) were performed in 8-week-old mice at 24 h intervals and BrdU staining was assessed 35 days later. Briefly, mice were anaesthetized with 2.5% avertin in saline, perfused with 20 U ml^{-1} heparin in PBS, and subsequently with 4% PFA in PBS. Brain were collected, immersion fixed overnight at 4 °C, infiltrated with 30% sucrose in PBS overnight at 4 °C and flash-frozen in tissue freezing medium. Sections of 17 μm were cut and air-dried overnight at room temperature. Subsequently, the sections were incubated in 2 N HCl at 37 °C for 30 min, rinsed for 10 min in 0.1 M borate buffer (pH 8.4) at room temperature and washed six times in PBS. To block endogenous peroxidase activity, sections were incubated for 10 min in 0.3% H_2O_2 in 10% methanol. Non-specific binding was blocked by 1 h incubation in blocking buffer (PBS containing 0.25% Triton-X and 5% donkey serum). Sections were incubated for 72 h at 4 °C with primary antibody (anti-BrdU) diluted in blocking buffer, then washed in PBS 0.1% Tween-20, and incubated for 2 h at room temperature in the dark with a biotin-conjugated secondary antibody diluted in blocking buffer. Sections were washed and 3,3′-diaminobenzidine (DAB) staining was performed using Vectastain Elite ABC-kit (Vector Laboratories) and DAB peroxidase (HRP) substrate kit (Vector Laboratories) following manufacturer's instructions. Afterwards, sections were counterstained using haematoxylin for 15 s, sealed with Vectamount (Vector Laboratories) and stored at room temperature. Images were acquired with a Zeiss slide scanner Axio Scan.Z1 using 20× objective and analysed with ImageJ. For quantitation, BrdU^{+} cells in granule cell layer and glomerular layer of the olfactory bulb were counted in two consecutive sections per animal and averaged per animal.

Immunocytochemistry. Cells were fixed for 10 min in 4% PFA/4% sucrose in PBS, washed with PBS, incubated for 10 min in 100 mM glycine, permeabilized for 5 min in 0.2% Triton-X100 in PBS, washed with PBS, and nonspecific binding was then blocked by 40 min incubation in blocking buffer (2% IgG-free BSA in PBS). Coverslips were incubated for 1 h at 22–24 °C with primary antibody diluted in blocking buffer, washed five times in PBS, and then incubated for 1 h at 22–24 °C in the dark with Hoechst stain and fluorophore-coupled donkey secondary antibody (identified above) diluted in blocking buffer. Coverslips were washed and mounted on slides with Fluoromount-G (SouthernBiotech) and stored at 4 °C. Images were acquired with a Zeiss LSM 710 confocal microscope using Plan-Apochromat 40× and 63× objectives.

CLARITY imaging. One cerebral hemisphere from a *Cx3cr1*$^{GFP/+}$ mouse was cleared using CLARITY protocols, essentially as described[34]. Rather than electrophoretic clearing, samples were incubated at 37 °C and passively cleared over 3 weeks by daily replacement of the clearing solution. A 1 mm^3 block of tissue adjacent to the lateral ventricle of *Cx3cr1*$^{GFP/+}$ mice, in the region containing the SVZ, was imaged using a Zeiss LSM 710 confocal microscope. Fiji software was used to assemble images.

Immunoblot. Cultured cells were washed with ice-cold DPBS and lysed on ice in 50 mM Tris-HCl (pH 7.5), 1 mM EGTA, 1 mM EDTA, 1% Triton-X100, 0.27 M sucrose, and protease and phosphatase inhibitors (Roche). Tissues were snap-frozen in liquid nitrogen before lysis. For immunoblot analysis, equal amounts of protein in LDS sample buffer (Invitrogen) were separated by electrophoresis through 4–12% Bis-Tris polyacrylamide gels (Novex, Life Technologies) and transferred to PVDF membranes (Millipore). For Axl immunoprecipitation, tissue lysates were precleared overnight at 4 °C with Protein G-Sepharose (Invitrogen). This was then removed and lysates were incubated for 2 h with 0.2 μg anti-Axl (M20) for 0.5 mg protein in cell lysate. Fresh Protein G–Sepharose was added for 2 h and immunoprecipitates were washed twice with 1 ml of lysis buffer containing 0.5 M NaCl and once with 1 ml of 50 mM Tris-HCl (pH 7.5). Immunoprecipitates were eluted in LDS buffer, separated by electrophoresis through polyacrylamide gels and transferred to PVDF membranes. Nonspecific binding was blocked with TBST (50 mM Tris-HCl (pH 7.5), 0.15 M NaCl and 0.1% Tween-20) containing 5% BSA, and membranes were incubated overnight at 4 °C with primary antibodies diluted in blocking buffer. Blots were then washed in TBST and incubated for 1 h at 22–24 °C with secondary horseradish peroxidase–conjugated antibodies in 5% skim milk in TBST. After repeating the washes, signal was detected with enhanced chemiluminescence reagent.

Reverse transcription (RT)–qPCR. Total cellular RNA was isolated with an RNeasy Mini Kit according to the manufacturer's instructions (Qiagen).

DNA was removed by on-column digestion with DNase (Qiagen). An RT Transcriptor First Strand cDNA Synthesis Kit (Roche) with anchored oligonucleotide (dT) primers (Roche) was used for reverse transcription. Quantitative PCR was run in a 384-well plate format on a ViiA 7 Real-Time PCR System (Applied Biosystems) with 2 × SYBR Green PCR Master Mix (Applied Biosystems). Primers are listed in Supplementary Table 1. Expression was analysed by the threshold cycle ($\Delta\Delta C_t$).

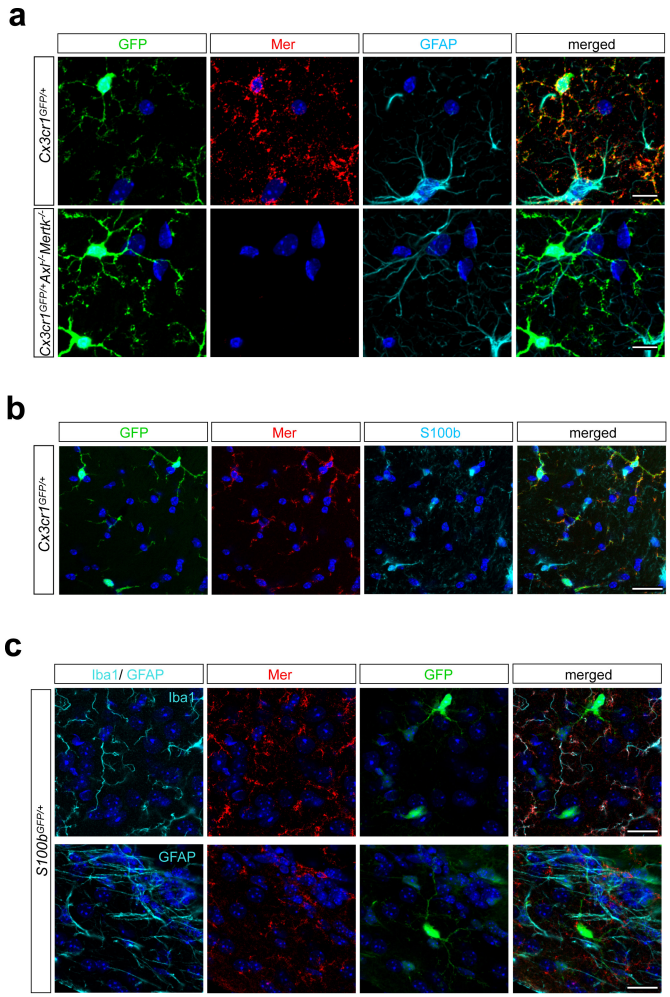
Microglia and astrocyte culture. Postnatal day 30 (P30) to P50 mice (*Cx3cr1*$^{GFP/+}$, *Axl*$^{-/-}$*Mertk*$^{-/-}$*Cx3cr1*$^{GFP/+}$) brains were dissociated using Neural Dissociation kit, Postnatal Neurons and the gentleMACS dissociator according to the manufacturer's instructions (Miltenyi). Single cell suspensions were resuspended in 30% Percoll in HBSS solution and centrifuged 15 min at 700 g to remove myelin. Cells were grown for 7 days in DMEM-F12 with 10% fetal bovine serum (FBS) and 1% penicillin/streptomycin before being processed for immunostaining or phagocytosis assay. Cytosine β-D-arabinofuranoside (Ara-C, 5 μM) was added after 5 days *in vitro* to limit fibroblast proliferation. When astrocytes were also isolated, microglia were first purified using C11b MicroBeads (Miltenyi) and grown for 7 days in DMEM-F12 with 10% FBS and 1% penicillin/streptomycin while astrocytes were grown for 10 days in MACS Neuro Medium with 2% MACS NeuroBrew-21 and 1% penicillin/streptomycin.

Phagocytosis assay. For the generation of apoptotic cells, thymocytes were isolated from 3- to 6-week-old mice, red blood cells were lysed with ACK buffer and remaining cells were incubated for 6 h in RPMI medium containing 5% FBS and 2 μM dexamethasone to induce apoptosis. This routinely resulted in 70% apoptotic and ≤5% necrotic cells. Apoptotic cells were then stained for 30 min with100 ng ml^{-1} pHrodo-s.e. (Invitrogen) as described previously[12,35,36]. Labelled cells were washed twice in PBS containing 1% BSA (to block remaining pHrodo-SE) and 1 mM EDTA (to remove any bound Gas6 and protein S) and once with DMEM. Apoptotic cells were then incubated for 10 min with recombinant mouse Gas6 or purified human protein S, added to microglia or astrocyte cultures at a ratio of 10:1 (apoptotic cells:phagocytes), and incubated for 1 h at 37 °C. Microglia or astrocytes were then briefly washed in DPBS, incubated for 10 min at 37 °C in trypsin (0.25%), and then placed on ice and detached by vigorous pipetting. Astrocytes were labelled using anti-ACSA2-APC antibody[37]. Phagocytosis was assessed by flow cytometry with post-acquisition data analysis with FlowJo software (TreeStar). pHrodo fluorescence was measured with excitation at 561 nm and emission filters for phycoerythrin (574–590 nm) on a LSR II (BD Biosciences) at the Flow Cytometry Core of the Salk Institute, as described previously[12]. Microglia were gated as GFP$^+$ cells and astrocytes were gated as APC$^+$ cells.

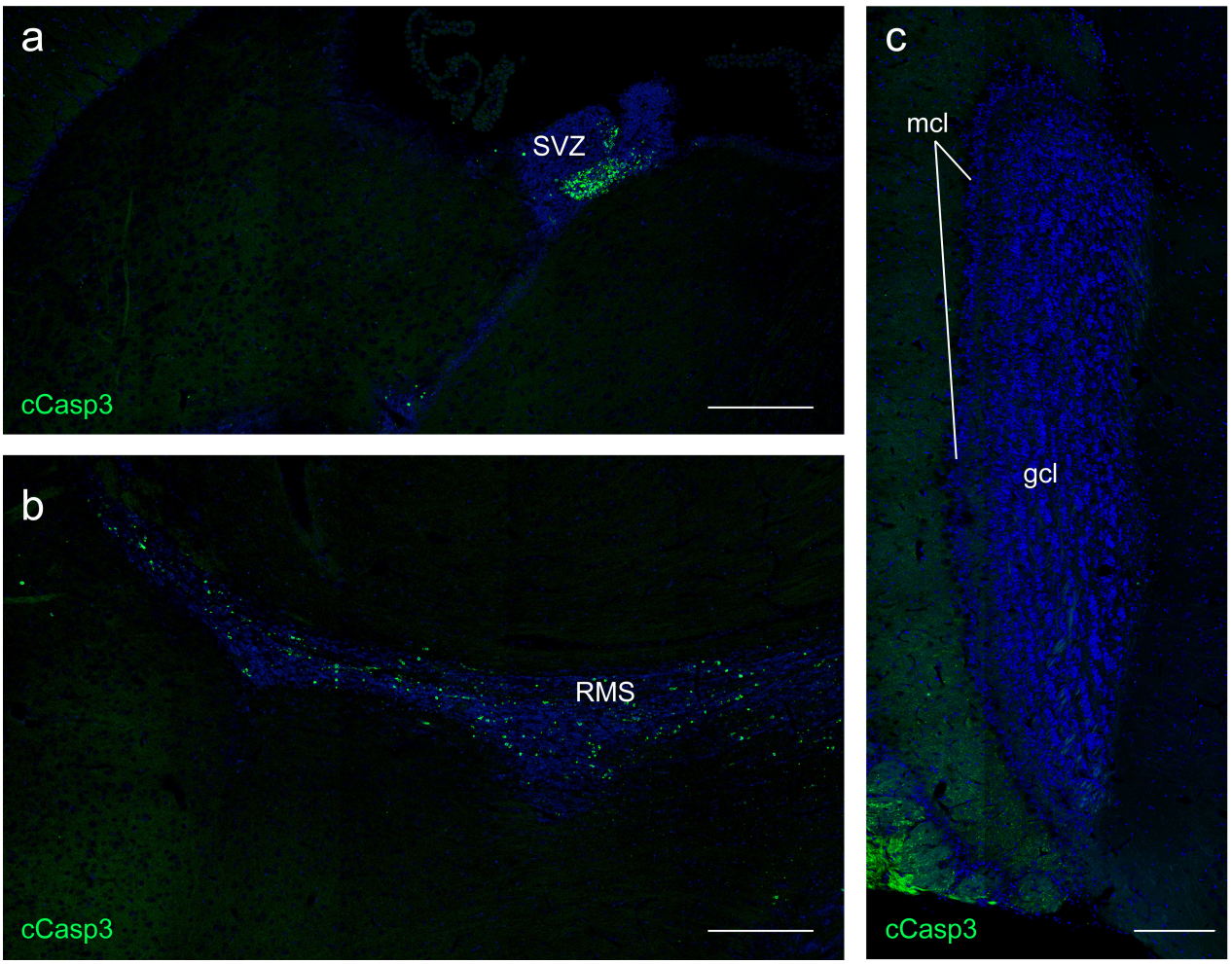
Two-photon imaging. Adult male mice (3–6 months old) were anaesthetized with isoflurane (1.5–2.5% in 100% oxygen at 0.8–1.0 l min^{-1}). Body temperature was kept at 36–37 °C, and hydration status was maintained using subcutaneous physiological saline injections (0.1 ml per 25 g body weight every 1–2 h). For head plate implantation, hair, skin and periosteum overlying the neocortex were removed. After cleaning exposed skull areas, a custom metal head plate was affixed to the skull using OptiBond (31514; Kerr) and dental acrylic (H00335; Coltene Whaledent), keeping the intended imaging area over somatosensory or visual cortex uncovered. A polished and reinforced thinned skull window (~2–3-mm diameter; ~20–50 μm remaining bone thickness) was then prepared, as described previously[38,39]. A movable objective microscope (Sutter Instrument) equipped with a pulsed femtosecond Ti:Sapphire laser (Chameleon Vision II or Ultra II, Coherent), two fluorescence detection channels (565DXCR dichroic, ET525/70M-2P and ET605/70M-2P emission filters, Chroma; H7422-40 GaAsP photomultiplier tubes, Hamamatsu), and a water immersion objective (LUMPlanFL N 40XW 0.8NA; Olympus) was used for two-photon imaging. Imaging was performed as described previously[28,39] using 920–940 nm centre excitation wavelengths. Average laser powers used for transcranial optical recordings depended on imaging depth (typically ~10–30 mW at ~150–200 μm depth from the pia). Images were typically acquired using a 6 Hz frame rate, 256 × 256 pixel resolution and a 5-frame average. Image stacks were acquired every 1.5–2 min for up to 5 h and typically contained 20–30 images per stack with 1 μm axial image spacing. Fields-of-view had a typical side length of 65–100 μm. Imaging settings were kept constant during time-lapse recordings. For quantitative image data analysis, ImageJ or Fiji software was used. First, maximum intensity images were produced from individual image stacks. Then, lateral image shifts in time-lapse recordings were corrected using a custom-written ImageJ alignment plugin based on the position shift of the peak in cross-correlation images, typically using the first projection image as the reference image. Structural dynamics of individual microglial cell processes was quantified manually using the MTtrackJ plugin in Fiji. Image analysis was done blind with respect to experimental condition. Videos were also created with Fiji.

Laser lesion. To target blood vessels for focal laser lesion, blood plasma was stained by tail vein injection of biocytin-TMR (2–5% in saline, T-12921, Life Technologies). Lesions were performed following a baseline recording period of 30–45 min, during which *z*-stacks were acquired as described above. To induce lesions, the Ti:Sapphire laser was transiently tuned to 800 nm and a confined area (8–15 μm diameter, ~1 μm axial extent) of a horizontally oriented cortical capillary at 150–220 μm depth was exposed to 70–130 mW for 10–30 s. Laser lesions caused extravasation of dye, indicating disruption of the blood-brain barrier. Following focal lesion, image stack acquisition was resumed using the same laser and recording parameters as during the baseline recording period. Although Supplementary Videos 4 and 5 run for only ~12 min (the time required for microglial processes to reach the lesion site), time-lapse recording of the same cortical volume continued for 2–4 h after the lesion.

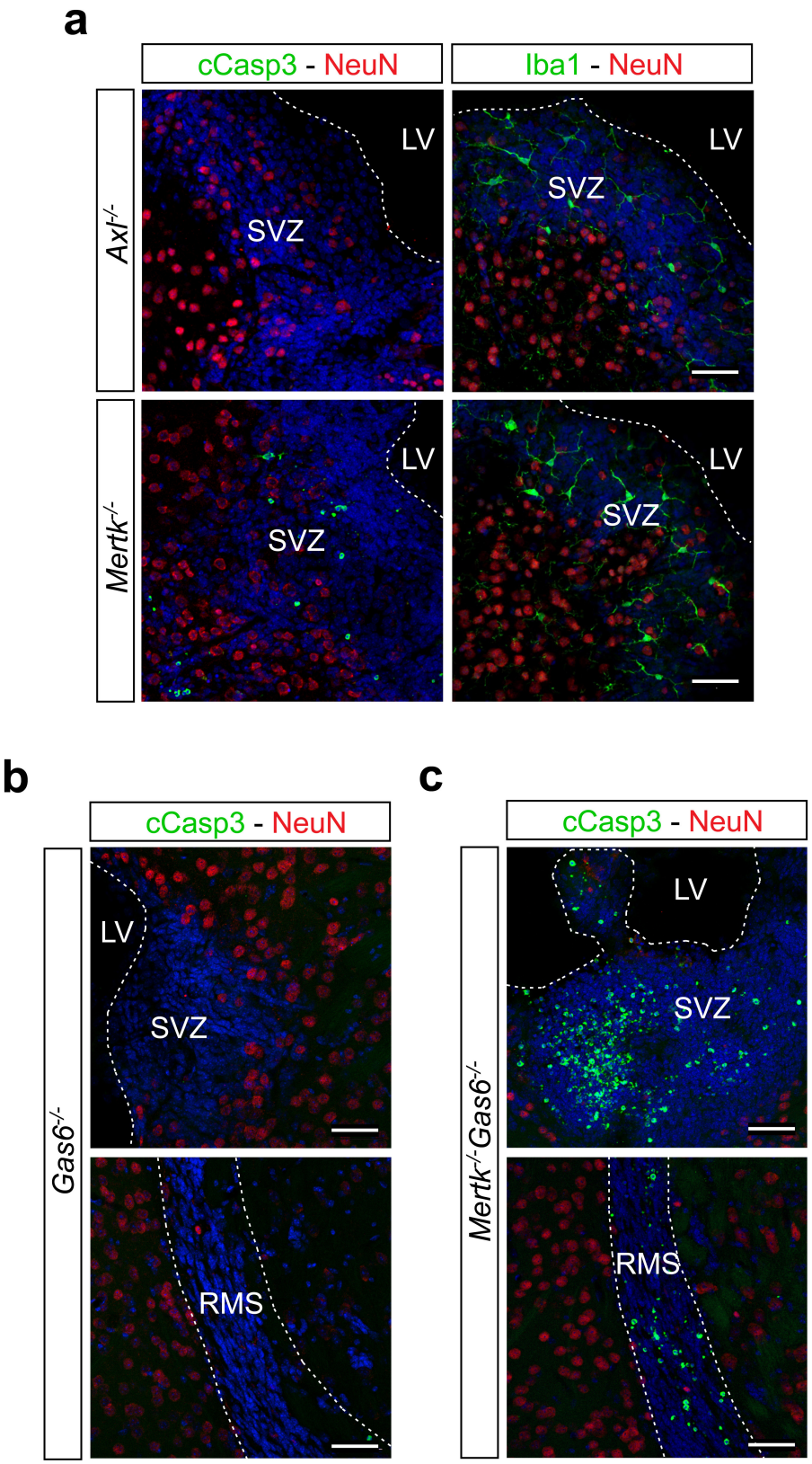
31. Lu, Q. *et al.* Tyro-3 family receptors are essential regulators of mammalian spermatogenesis. *Nature* **398,** 723–728 (1999).
32. Angelillo-Scherrer, A. *et al.* Deficiency or inhibition of Gas6 causes platelet dysfunction and protects mice against thrombosis. *Nature Med.* **7,** 215–221 (2001).
33. Vives, V., Alonso, G., Solal, A. C., Joubert, D. & Legraverend, C. Visualization of S100B-positive neurons and glia in the central nervous system of EGFP transgenic mice. *J. Comp. Neurol.* **457,** 404–419 (2003).
34. Chung, K. *et al.* Structural and molecular interrogation of intact biological systems. *Nature* **497,** 332–337 (2013).
35. Miksa, M., Komura, H., Wu, R., Shah, K. G. & Wang, P. A novel method to determine the engulfment of apoptotic cells by macrophages using pHrodo succinimidyl ester. *J. Immunol. Methods* **342,** 71–77 (2009).
36. Dransfield, I., Zagorska, A., Lew, E. D., Michail, K. & Lemke, G. Mer receptor tyrosine kinase mediates both tethering and phagocytosis of apoptotic cells. *Cell Death Dis.* **6,** e1646 (2015).
37. Sharma, K. *et al.* Cell type- and brain region-resolved mouse brain proteome. *Nature Neurosci.* **18,** 1819–1831 (2015).
38. Drew, P. J. *et al.* Chronic optical access through a polished and reinforced thinned skull. *Nature Methods* **7,** 981–984 (2010).
39. Knowland, D. *et al.* Stepwise recruitment of transcellular and paracellular pathways underlies blood-brain barrier breakdown in stroke. *Neuron* **82,** 603–617 (2014).



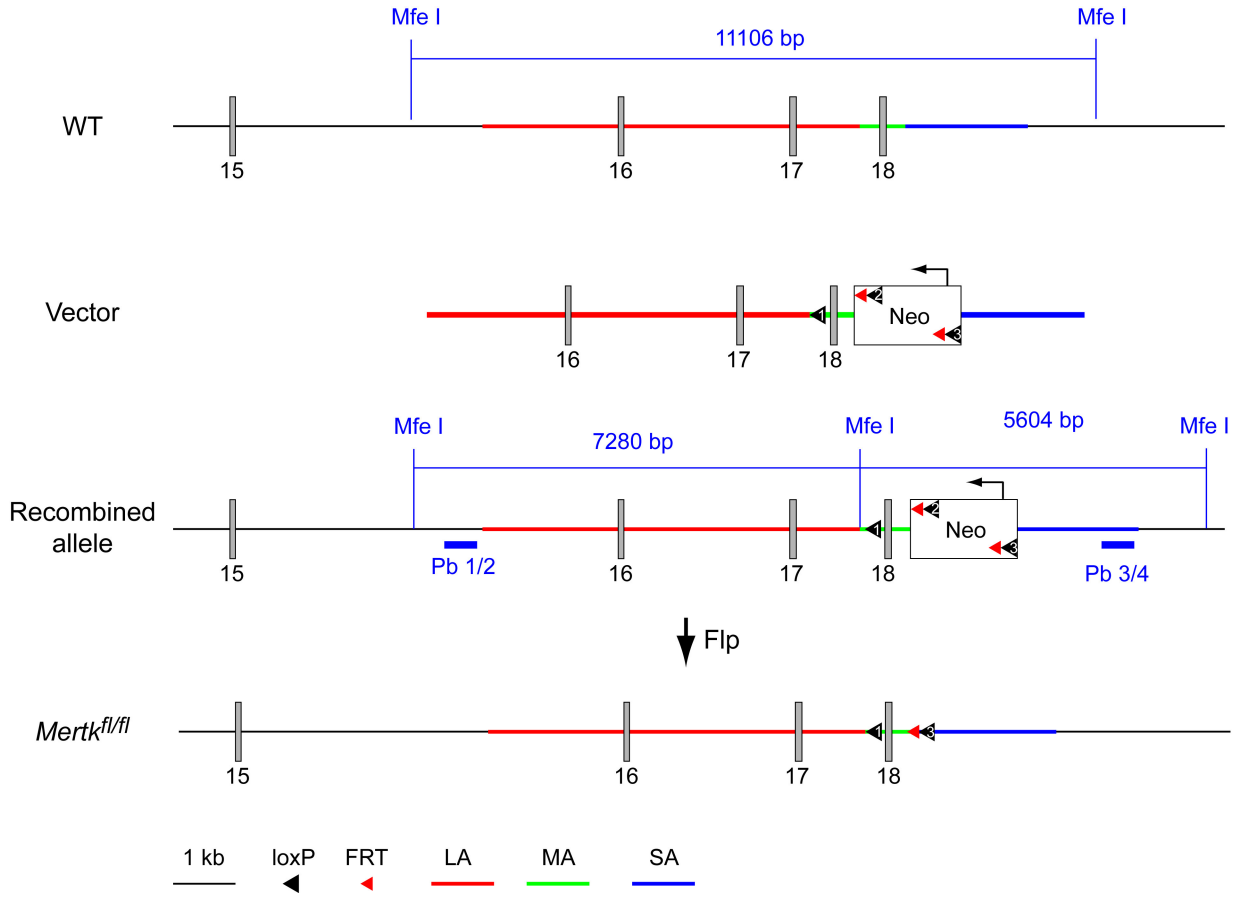
Extended Data Figure 1 | Mer is expressed by microglia. a, Brain (hippocampus) sections from *Cx3cr1*$^{GFP/+}$ mice that were wild-type (top row) or *Axl*$^{-/-}$*Mertk*$^{-/-}$ (bottom row) were visualized by confocal microscopy for GFP (1st column), anti-Mer (red, 2nd column), or anti-GFAP (cyan, 3rd column) immunoreactivity; 4th column, merged images. Scale bars, 10 μm. Axl immunostaining signal is too low to be visualized in unactivated microglia (not shown; but see Fig. 4d). **b**, Mer expression does not co-localize with S100b^{+} cells. Immunostaining of *Cx3cr1*$^{GFP/+}$ brain sections with anti-Mer (red, 2nd panel) and anti-S100b (cyan, 3rd panel); 4th panel, merged images. **c**, Mer co-localizes with Iba1, but not GFAP or GFP in *S100b*$^{GFP/+}$ mice. Brain sections were visualized by confocal microscopy for anti-Iba1 (top) or anti-Gfap (bottom) (both cyan, 1st column), anti-Mer (red, 2nd column), or GFP (green, 3rd column) immunoreactivity; 4th column, merged images. Scale bars (**b**, **c**), 20 μm. Representative images from analyses performed in $n = 2$ mice (**a**–**c**).



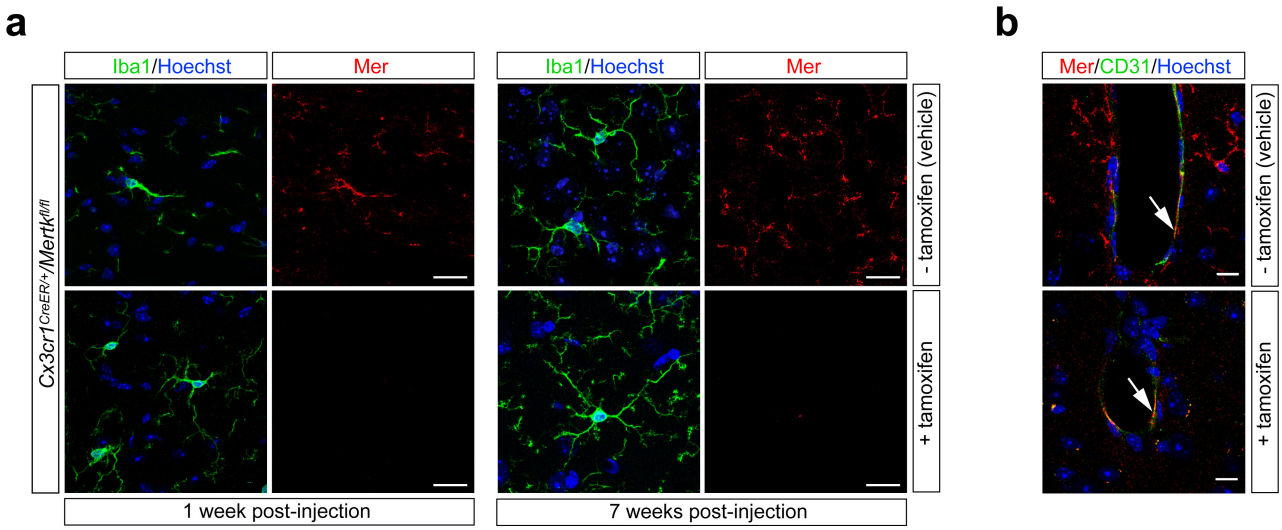
Extended Data Figure 2 | Accumulation of apoptotic cells is confined to neurogenic and derivative migratory regions of the *Axl*$^{-/-}$*Mertk*$^{-/-}$ CNS. a, A low power tiled image of a section through the *Axl*$^{-/-}$*Mertk*$^{-/-}$ subventricular zone and surrounding brain tissue, stained for cCasp3, illustrates that apoptotic cells are confined within the SVZ. **b**, A low power tiled image of a section through the *Axl*$^{-/-}$*Mertk*$^{-/-}$ rostral migratory stream (RMS) and surrounding brain tissue illustrates that cCasp3^{+} apoptotic cells are confined within the RMS. **c**, A low power tiled image of the granule cell and mitral cell layers (gcl and mcl, respectively) of the *Axl*$^{-/-}$*Mertk*$^{-/-}$ olfactory bulb, stained for cCasp3, illustrates that there are no apoptotic cells detected in the double mutant bulb. Scale bars (**a**–**c**), 200 μm. Representative images from analyses performed in $n = 3$ mice (**a**–**c**).



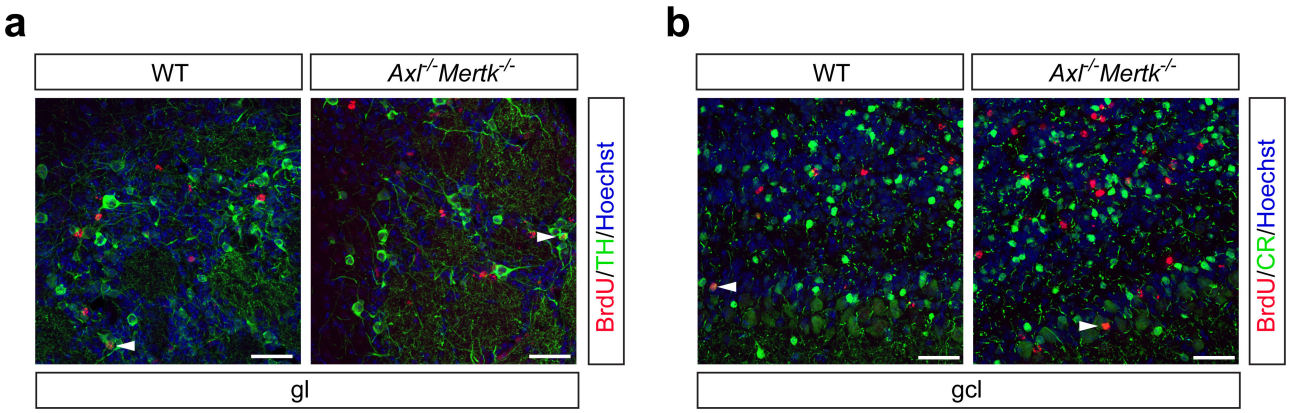
Extended Data Figure 3 | Mer is the principal microglial TAM receptor required for phagocytosis of apoptotic cells in the SVZ. a, Sections of the SVZ from $Axl^{-/-}$ (top row) and $Mertk^{-/-}$ (bottom row) mice immunostained for cCasp3 and NeuN (green and red, respectively; left panels), or Iba1 and NeuN (green and red, respectively; right panels) reveal the accumulation of cCasp3^{+} apoptotic cells only in the $Mertk^{-/-}$ SVZ. **b**, Sections of the SVZ (top) and RMS (bottom) of $Gas6^{-/-}$ mice, illustrating no accumulation of apoptotic cells (similar to both wild type and $Axl^{-/-}$). **c**, Sections of the SVZ (top) and RMS (bottom) of $Mertk^{-/-}Gas6^{-/-}$ mice, illustrating a massive accumulation of aoptotic cells, similar to that seen in $Axl^{-/-}Mertk^{-/-}$ mice. Scale bars, 50 μm. See main text for quantification. Representative images from analyses performed in $n = 2$ mice for $Gas6^{-/-}$ and $Mertk^{-/-}Gas6^{-/-}$, and $n = 3$ mice for $Axl^{-/-}$ and $Mertk^{-/-}$.



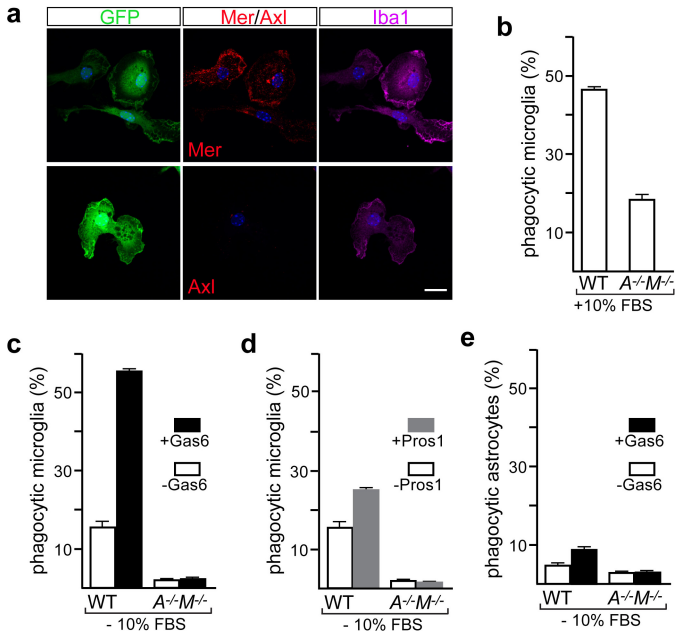
Extended Data Figure 4 | Conditional *Mertk* knockouts. The knockout strategy targets exon 18 of the wild-type mouse *Mertk* gene, which encodes residues W779–L824 of the tyrosine kinase domain (1st line). Deletion of this exon leads to a functional and null protein (see Methods, and Extended Data Fig. 5). The targeting vector (2nd line) had a PGK-Neo cassette for selection in embryonic stem (ES) cells, and contained loxP and FRT sites, recognized by Cre and Flp recombinases, respectively, at the indicated positions. Five ES cell lines with homologous recombination at the *Mertk* locus were identified by Southern blots of MfeI-digested DNA, using the indicated Pb 1/2 (external) and Pb 3/4 (internal) probes (3rd line). Introduction of Flp recombinase, achieved by crossing high percentage chimaeras (obtained from blastocyst injection of these ES cells) to C57Bl/6 FLP mice, removed the Neo cassette, leaving exon 18 flanked by loxP sites (4th line). Cre-mediated recombination at these loxP sites deletes exon 18. *Mertk*$^{fl/fl}$ mice, together with PCR-based protocols for their genotyping, are available upon request from the Rothlin laboratory (contact C.V.R.). See Methods for further information.



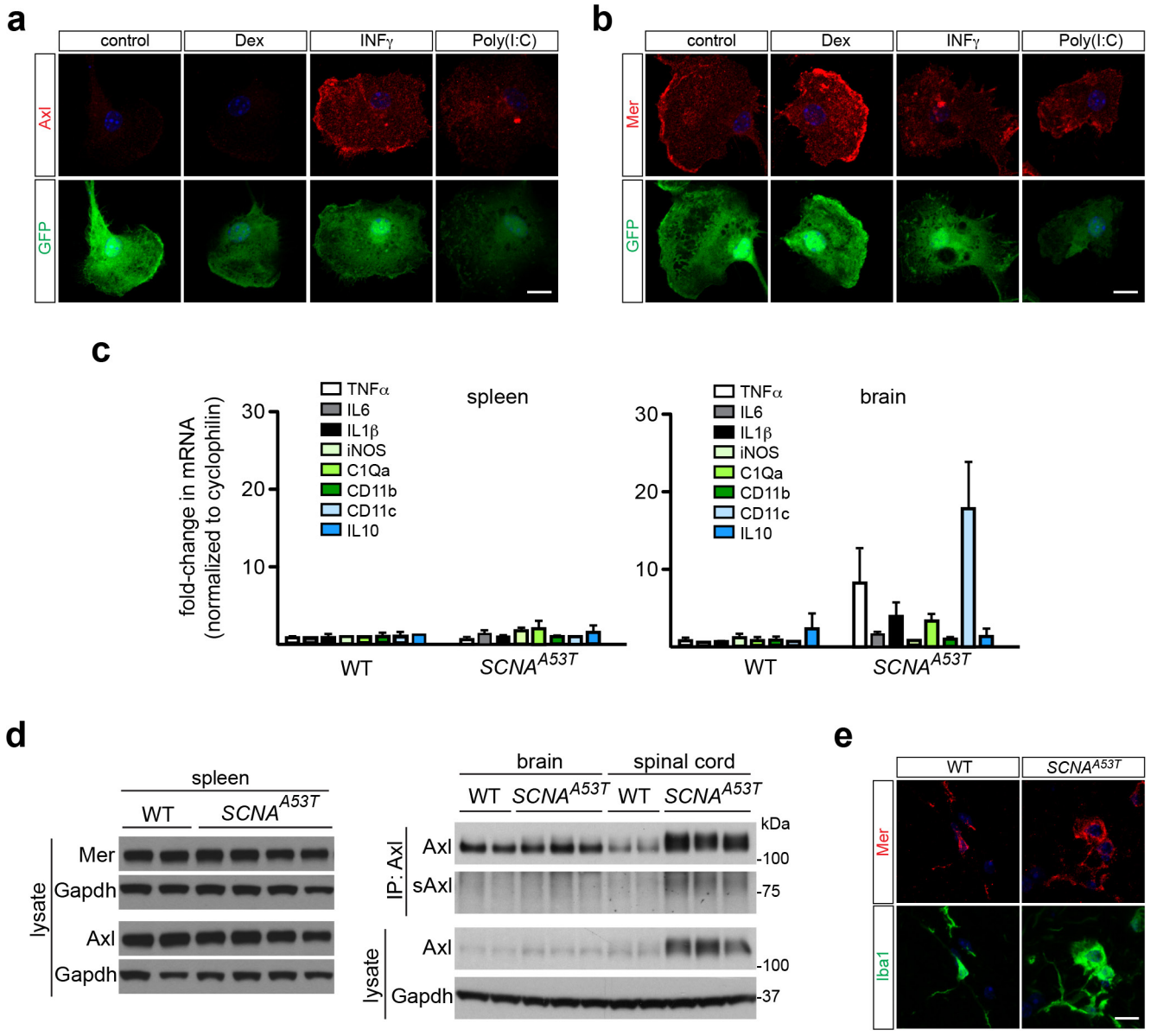
Extended Data Figure 5 | Persistence of microglial-specific Mer ablation following tamoxifen injection of *Cx3cr1*$^{CreER/+}$*Mertk*$^{fl/fl}$ mice. **a**, Mice were injected intraperitoneally with oil vehicle alone (−tamoxifen, top row) or with tamoxifen (+tamoxifen, bottom row) (see Methods), and brain sections were immunostained for Mer protein expression (red panels in 2nd, 4th columns) in Iba1$^+$ microglia (green panels in 1st, 3rd columns) at 1 week (left four panels) and 7 weeks (right four panels) after injection. Sections counter-stained with Hoechst 33258 to visualize nuclei (blue). **b**, Brain sections containing a brain capillary 7 weeks after injection of vehicle (top) or tamoxifen (bottom), showing that although Mer expression in microglia is eliminated upon tamoxifen-mediated *Cx3cr1*-restricted induction of Cre activity, Mer expression in CD31$^+$ microvascular endothelial cells (arrows) is maintained. Representative images of $n = 2$ mice per time point.



Extended Data Figure 6 | Identity of immigrant BrdU$^+$ cells in the olfactory bulb. a, A group of the BrdU$^+$ cells in the glomerular layer (gl), visualized 35 days after injection of BrdU (red) and presumed immigrant descendants of SVZ cells in S phase at the time of injection, are also positive for tyrosine hydroxylase (TH, green) in both wild-type (left panel) and *Axl*$^{-/-}$*Mertk*$^{-/-}$ (right panel) mice. Arrowheads are examples of TH$^+$BrdU$^+$ cells. **b**, Similar comparative granule cell layer (gcl) sections stained with anti-BrdU (red) and calretinin (CR, green). Arrowheads are examples of CR$^+$BrdU$^+$ cells. Sections were co-stained with Hoechst 33258 to visualize nuclei. Scale bars, 50 μm. Representative images of $n = 2$ per genotype.



Extended Data Figure 7 | Both Gas6 and Pros1 drive microglial phagocytosis of apoptotic cells *in vitro*. **a**, Cultured microglia express Mer but little or no Axl under basal conditions. Microglia were cultured from wild-type *Cx3cr1*$^{GFP/+}$ mice, visualized for GFP (1st column), and immunostained for Iba1 (3rd column), Mer (2nd column, top), and Axl (2nd column, bottom). Scale bar, 10 μm. **b**–**d**, *In vitro* pHrodo-based assay of phagocytosis of apoptotic cells by microglia (see Methods). **b**, In serum-containing medium (10% FBS), wild-type microglia are effective phagocytes; mean phagocytic activity is substantially reduced in *Axl*$^{-/-}$*Mertk*$^{-/-}$ (*A*$^{-/-}$*M*$^{-/-}$) microglia. **c**, **d**, Both purified Gas6 (**c**) and purified Pros1 (**d**) stimulate AC phagocytosis by cultured microglia in serum-free medium, and this stimulation is entirely TAM-dependent. **e**, The phagocytic activity of cultured astroglia prepared from *Cx3cr1*$^{GFP/+}$ mice that were either wild-type or *Axl*$^{-/-}$*Mertk*$^{-/-}$ was measured in the same pHrodo-based assay in serum-free medium ± Gas6. For this FACS-based assay, astrocytes were gated using an astrocyte-specific surface antigen-2 (ACSA-2) antibody (see Methods). Bar graphs represent mean phagocytic activity (± s.em.); $n = 2$ replicates from 2 mice per genotype (**b**–**d**), and 2 replicates from 4 mice per genotype (**e**).



Extended Data Figure 8 | Regulation of microglial Axl by neuroinflammation. a, **b**, Axl (**a**) and Mer (**b**) regulation in purified (GFP^+) cultured microglia by the tolerogenic stimulus dexamethasone (Dex) and the two proinflammatory stimuli IFNγ and poly(I:C), as assessed by immunostaining. Axl expression (**a**) is very low in the absence of an added stimulus, is not elevated by Dex, but is strongly upregulated by both IFNγ and poly(I:C). In contrast, Mer expression (**b**) is readily detected in the absence of an added stimulus, is further elevated by Dex, but is modestly suppressed by both IFNγ and poly(I:C). Scale bar, 10 μm. **c**, In contrast to the spinal cord (see Fig. 4a), there is no upregulation of the indicated inflammatory mediator/marker mRNAs (mean expression ± s.e.m.) in the spleens, and only modest upregulation in the brains, of $SNCA^{A53T}$ mice at 8–10 months of age. $n = 3$ mice for each genotype. **d**, Western blot analysis of spleen (left blots) and brain and spinal cord (right blots) extracts from two different wild-type mice and four or three different $SNCA^{A53T}$ mice at 9–10 months, for the indicated proteins, with Gapdh as a loading control. Note that soluble Axl ectodomain (sAxl) is upregulated in the $SNCA^{A53T}$ spinal cord concomitantly with Axl. **e**, Although Axl is strongly upregulated in $Iba1^+$ microglia in the $SNCA^{A53T}$ spinal cord (see Fig. 4d), no upregulation of Mer is observed in these same cells. Scale bar, 10 μm. $n = 2$ wild-type and 3 $SNCA^{A53T}$ mice.

LETTER

doi:10.1038/nature17403

The necrosome promotes pancreatic oncogenesis via CXCL1 and Mincle-induced immune suppression

Lena Seifert[1]*, Gregor Werba[1]*, Shaun Tiwari[1], Nancy Ngoc Giao Ly[1], Sara Alothman[1], Dalia Alqunaibit[1], Antonina Avanzi[1], Rocky Barilla[1], Donnele Daley[1], Stephanie H. Greco[1], Alejandro Torres-Hernandez[1], Matthew Pergamo[2], Atsuo Ochi[1], Constantinos P. Zambirinis[1], Mridul Pansari[1], Mauricio Rendon[1], Daniel Tippens[1], Mautin Hundeyin[1], Vishnu R. Mani[1], Cristina Hajdu[3], Dannielle Engle[4] & George Miller[1,2]

Neoplastic pancreatic epithelial cells are believed to die through caspase 8-dependent apoptotic cell death, and chemotherapy is thought to promote tumour apoptosis[1]. Conversely, cancer cells often disrupt apoptosis to survive[2,3]. Another type of programmed cell death is necroptosis (programmed necrosis), but its role in pancreatic ductal adenocarcinoma (PDA) is unclear. There are many potential inducers of necroptosis in PDA, including ligation of tumour necrosis factor receptor 1 (TNFR1), CD95, TNF-related apoptosis-inducing ligand (TRAIL) receptors, Toll-like receptors, reactive oxygen species, and chemotherapeutic drugs[4,5]. Here we report that the principal components of the necrosome, receptor-interacting protein (RIP)1 and RIP3, are highly expressed in PDA and are further upregulated by the chemotherapy drug gemcitabine. Blockade of the necrosome *in vitro* promoted cancer cell proliferation and induced an aggressive oncogenic phenotype. By contrast, *in vivo* deletion of RIP3 or inhibition of RIP1 protected against oncogenic progression in mice and was associated with the development of a highly immunogenic myeloid and T cell infiltrate. The immune-suppressive tumour microenvironment associated with intact RIP1/RIP3 signalling depended in part on necroptosis-induced expression of the chemokine attractant CXCL1, and CXCL1 blockade protected against PDA. Moreover, cytoplasmic SAP130 (a subunit of the histone deacetylase complex) was expressed in PDA in a RIP1/RIP3-dependent manner, and Mincle—its cognate receptor—was upregulated in tumour-infiltrating myeloid cells. Ligation of Mincle by SAP130 promoted oncogenesis, whereas deletion of Mincle protected against oncogenesis and phenocopied the immunogenic reprogramming of the tumour microenvironment that was induced by RIP3 deletion. Cellular depletion suggested that whereas inhibitory macrophages promote tumorigenesis in PDA, they lose their immune-suppressive effects when RIP3 or Mincle is deleted. Accordingly, T cells, which are not protective against PDA progression in mice with intact RIP3 or Mincle signalling, are reprogrammed into indispensable mediators of anti-tumour immunity in the absence of RIP3 or Mincle. Our work describes parallel networks of necroptosis-induced CXCL1 and Mincle signalling that promote macrophage-induced adaptive immune suppression and thereby enable PDA progression.

We found that RIP1 and RIP3 are highly expressed in human PDA (Fig. 1a, b). Western blotting confirmed that expression of RIP1 and RIP3 was higher in human PDA than in the surrounding normal pancreas (Fig. 1c). Similarly, FADD (which complexes with RIP1/RIP3 to form the necrosome), MLKL (a downstream mediator of necroptosis), and caspase 8 (a principal driver of apoptosis) were upregulated in PDA[5] (Fig. 1c). Immunofluorescence microscopy showed that RIP1 and RIP3 co-localized in human (Fig. 1d) and mouse (Fig. 1e) PDA, consistent with necrosome complex formation. To test whether necrosome formation was induced by chemotherapy, we treated PDA-bearing mice with gemcitabine. Gemcitabine treatment increased expression of RIP1 and RIP3 in PDA *in vivo* (Fig. 1f, g). Similarly, gemcitabine increased expression of *Rip1* and *Rip3* (also known as *Ripk1* and *Ripk3*, respectively) and RIP1–RIP3 co-localization *in vitro* in PDA cells (Fig. 1h, i). Chemotherapy also induced components of the necrosome in human PDA cells, whereas MLKL inhibition prevented chemotherapy-induced cell death in human PDA cells (Fig. 1j, k).

As necroptosis is a pro-inflammatory process, we postulated that it would support peri-tumoral inflammation[6]. We found that CXCL1 is one of the most highly expressed chemokines in mouse PDA (Fig. 2a). Similarly, CXCL1 was robustly expressed in human PDA (Fig. 2b–d). Gemcitabine upregulated the expression of CXCL1 in PDA in mice (Fig. 2e), whereas RIP3 deletion reduced CXCL1 expression *in vivo* (Fig. 2f, g) and *in vitro* (Fig. 2h). High *RIP3* expression also correlated with high *CXCL1* expression in a human PDA RNA sequencing (RNA-seq) database (Fig. 2i). Furthermore, upregulation of CXCL1 by gemcitabine was reduced by RIP3 deletion *in vivo* (Fig. 2j) and by RIP1 or RIP3 inhibition *in vitro* (Fig. 2k). Collectively, these data are consistent with necrosome-dependent upregulation of CXCL1 in PDA.

We studied the effects of RIP3 deletion on the properties of *in vitro* cultured $Kras^{G12D}$-transformed pancreatic ductal epithelial cells ($Kras^{G12D}$ PDEC). As expected, RIP3 deletion increased the proliferative rate of $Kras^{G12D}$ PDEC *in vitro* (Extended Data Fig. 1a). Moreover, $Kras^{G12D};Rip3^{-/-}$ PDEC exhibited a distinct phenotype including loss of CDK4 and elevated expression of Bcl-xL and c-Myc (Extended Data Fig. 1b), which are associated with aggressive tumour biology in PDA[7–11].

As RIP3 deletion increased the proliferation of PDA cells, we postulated that blockade of necroptosis *in vivo* would accelerate tumorigenesis. To test this theory, we compared the rate of oncogenic progression in $p48^{Cre};Kras^{G12D};Rip3^{+/+}$ and $p48^{Cre};Kras^{G12D};Rip3^{-/-}$ pancreases, which develop pancreatic neoplasia endogenously by expressing mutant *Kras* in the progenitor cells of the pancreas. Contrary to our hypothesis and belying our *in vitro* findings, RIP3 deletion protected against oncogenesis. $p48^{Cre};Kras^{G12D};Rip3^{-/-}$ pancreases showed a diminished rate of acinar replacement by dysplastic ducts, slower PanIN (pancreatic intraepithelial neoplasia) progression, and reduced fibro-inflammatory changes compared with $p48^{Cre};Kras^{G12D};Rip3^{+/+}$ pancreases (Fig. 3a and Extended Data Fig. 1c). Accordingly, aged-matched $p48^{Cre};Kras^{G12D};Rip3^{-/-}$ pancreases weighed less than controls and RIP3 deletion extended survival (Fig. 3b, c). The rate of proliferation was similar in $p48^{Cre};Kras^{G12D};Rip3^{+/+}$ and $p48^{Cre};Kras^{G12D};Rip3^{-/-}$ pancreatic epithelial cells *in vivo* (Fig. 3d). To test whether abrogation of RIP1 signalling also protected against PDA, we treated 6-week-old

[1]S. Arthur Localio Laboratory, Department of Surgery, New York University School of Medicine, 550 First Avenue, New York, New York 10016, USA. [2]Department of Cell Biology, New York University School of Medicine, 550 First Avenue, New York, New York 10016, USA. [3]Department of Pathology, New York University School of Medicine, 550 First Avenue, New York, New York 10016, USA. [4]Cold Spring Harbor Laboratories, Cold Spring Harbor, New York 11724, USA.
*These authors contributed equally to this work.

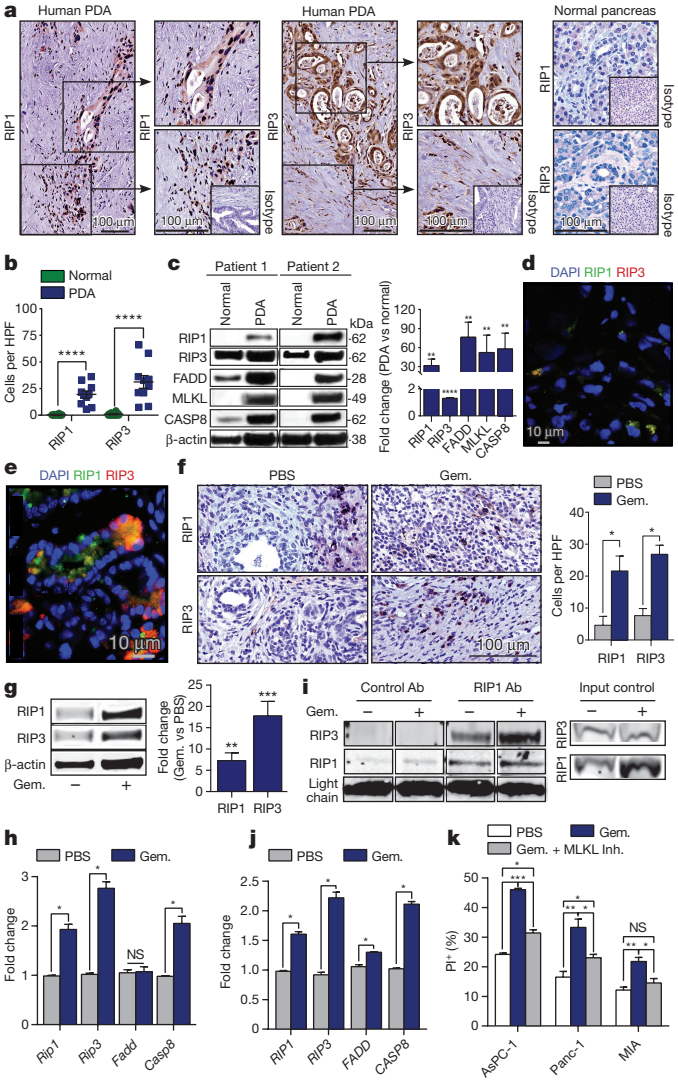


Figure 1 | RIP1 and RIP3 expression in PDA. a, **b**, Paraffin-embedded sections of human PDA and surrounding normal pancreas tissue from ten patients with PDA were tested for expression of RIP1 and RIP3. Representative images (**a**) and summary data (**b**) are shown. The magnification of isotype control slides are 4.5 times smaller than experimental slides. HPF, high-power field. **c**, Human PDA specimens and adjacent normal human pancreas tissue were tested for expression of RIP1, RIP3, FADD, MLKL, and cleaved caspase 8 (CASP8) by western blotting. Representative data from two patients and density analysis from four patients are shown. **d**, Frozen human PDA specimens were tested for RIP1 and RIP3 co-expression by immunofluorescence microscopy. A representative image is shown. Nuclei counterstained with 4′,6-diamidino-2-phenylindole (DAPI). **e**, Frozen sections of pancreas from 6-month-old $p48^{Cre};Kras^{G12D}$ (KC) mice, which express mutant *Kras*, were tested for RIP1 and RIP3 co-localization. **f**, $Pdx1^{Cre};Kras^{G12D};Tp53^{R172H}$ (KPC) mice, which express mutant *Kras* and *p53*, were serially treated with gemcitabine (Gem.) or PBS and tested for expression of RIP1 and RIP3 by IHC ($n = 3$ per group). **g**, Orthotopically implanted KPC-derived tumours were removed from gemcitabine- or PBS-treated mice and tested for RIP1 and RIP3 expression by western blotting. Representative data and averages of quadruplicates based on density analysis are shown. **h**, KPC-derived tumour cells were treated with gemcitabine or PBS *in vitro* in triplicate and tested for gene expression by quantitative PCR (qPCR). **i**, Lysate from $Pdx1^{Cre};Kras^{G12D};Tp53^{R172H}$ tumour cells that had been treated with PBS or gemcitabine was immunoprecipitated with control or anti-RIP1 antibodies (Ab) and tested for expression of RIP1 and RIP3. Input controls were also probed. **j**, Human AsPC-1 cells were treated with PBS or gemcitabine in triplicate and tested for gene expression by qPCR. **k**, AsPC-1, PANC1, and MIA PaCa-2 cells were cultured with PBS, gemcitabine, or gemcitabine and an MLKL inhibitor in quadruplicate. Cellular viability was determined at 24 h using propidium iodide (PI) staining. Graphs show mean ± s.e.m. n.s., not significant; $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, $****P < 0.0001$ (unpaired *t*-test). For gel source data, see Supplementary Fig. 1.

KC mice for 8 weeks with the RIP1 inhibitor Nec-1s. Nec-1s treatment protected against oncogenic progression, as assessed by pancreas weight and histology (Fig. 3e, f).

As necroptosis can increase inflammation[12], we postulated that RIP3 deletion protects against tumour progression by enhancing peri-tumoral immunogenicity. RIP3 deletion diminished infiltration by tumour-associated macrophages (TAMs; Extended Data Fig. 2a). Conversely, the fractions of T cells and B cells were increased in $p48^{Cre};Kras^{G12D};Rip3^{-/-}$ pancreases (Extended Data Fig. 2b, c). Analysis of the myeloid compartment showed a decreased fraction of myeloid-derived suppressor cells (MDSC) and dendritic cells in $p48^{Cre};Kras^{G12D};Rip3^{-/-}$ pancreases (Extended Data Fig. 2d, e). Furthermore, consistent with our immunohistochemical data, the number of bulk tumour-infiltrating TAMs and their M2-like $Arg1^{+}CD206^{+}$ subset were reduced by RIP3 deletion (Extended Data Fig. 2f–h). Macrophage expression of programmed death ligand 1 (PD-L1) was also reduced by RIP3 deletion (data not shown). Collectively, these data suggest that RIP3 deletion increases lymphocyte infiltration in PDA and reduces infiltration by immunosuppressive subsets of myeloid cells. Similarly, in human PDA, high *RIP1–RIP3* co-expression correlated with elevated expression of the myeloid cell marker *CD11b* (also known as *ITGAM*) (Extended Data Fig. 2i).

To determine whether deletion of RIP3 in the epithelial compartment alone is sufficient to protect against oncogenesis, we challenged wild-type mice with an orthotopic injection of either $Kras^{G12D};Rip3^{+/+}$ PDEC or $Kras^{G12D};Rip3^{-/-}$ PDEC. Similar to our findings using pan-RIP3 deletion, $Kras^{G12D};Rip3^{-/-}$ tumours grew more slowly than $Kras^{G12D};Rip3^{+/+}$ tumours (Fig. 4a), suggesting that RIP3 blockade in the epithelial compartment alone protects against PDA progression.

As inflammatory cells in the PDA tumour microenvironment (TME) express the components of the necrosome (Figs 1a and 4b, c), we investigated whether RIP3 deletion in the extra-epithelial compartment would similarly mitigate PDA progression. Wild-type and $Rip3^{-/-}$ mice were challenged with orthotopic intra-pancreatic injection of $Kras^{G12D}$ PDEC or $Pdx1^{Cre};Kras^{G12D};Tp53^{R172H}$ (KPC-derived) PDA cells, which express both mutant *Kras* and *p53* (also known as *Trp53*), and tumour size was measured 3 weeks later. $Rip3^{-/-}$ mice developed smaller $Kras^{G12D}$ (not shown) and $Pdx1^{Cre};Kras^{G12D};Tp53^{R172H}$ tumours (Fig. 4d) than did wild-type mice, consistent with the idea that blockade of necroptosis in the extra-epithelial compartment alone can protect against PDA.

To determine whether deletion of RIP3 in the extra-epithelial compartment similarly bolsters peri-tumoral immunogenicity, we analysed the inflammatory infiltrate in orthotopic KPC tumours in wild-type and $Rip3^{-/-}$ mice. RIP3 deletion resulted in elevated T cell and B cell infiltrates (Fig. 4e, f); peri-tumoral T cells expressed less IL-10 and PD-1 and more CD44, and included a lower fraction of T regulatory (T_{reg}) cells, in $Rip3^{-/-}$ mice than in wild-type mice (Fig. 4g, h; Extended Data Fig. 3a–d). Analysis of the myeloid compartment again revealed a reduction in the fraction of peri-tumoral MDSC (Fig. 4i) and TAMs (Fig. 4j), with a shift towards an M1-like phenotype (Fig. 4k, l) and reduced PD-L1 expression (Fig. 4m). These data appear to conflict with the recent finding that RIP1 signalling can enhance $CD8^{+}$ T cell cross-priming[13]. However, the effects in PDA may be unique to the immunological milieu of the pancreatic TME. Accordingly, RIP3 deletion was not protective against B16 melanoma or subcutaneously implanted KPC cells (Extended Data Fig. 3e, f).

As CXCL1 expression in PDA depends on the necrosome (Fig. 2) and we found that CXCR2 is widely expressed on peri-tumoral leukocytes (Extended Data Fig. 4a, b), we postulated that CXCL1 could mediate the

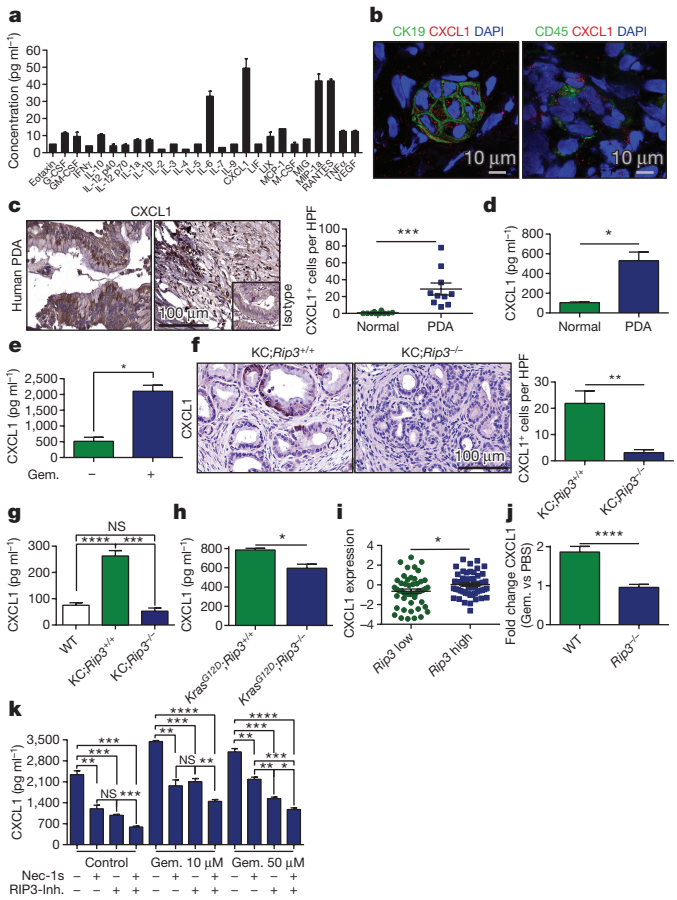


Figure 2 | CXCL1 is expressed in PDA in a RIP1/3-dependent manner. **a**, Single-cell suspensions of pancreas cells from 6-month-old KC mice were cultured for 24 h. Supernatant was tested for expression of inflammatory mediators. Averages of biological duplicates are shown. **b**, Human PDA tumours were tested for co-expression of CXCL1 and CK19 or CXCL1 and CD45 by confocal microscopy. Representative images are shown. **c**, Paraffin-embedded PDA sections from ten patients were tested for expression of CXCL1 by IHC and compared with surrounding normal pancreas. Representative ductal and stromal areas of PDA tumours and quantitative data are shown. The magnification of the isotype control insert is 4.5 times smaller than experimental slides. **d**, CXCL1 levels in tissue homogenate from three human PDA specimens were tested by ELISA and compared with surrounding normal pancreas. **e**, Homogenized KPC tumours from PBS- or gemcitabine-treated mice were tested for CXCL1 by ELISA. Experiments were performed in biological duplicates. **f**, Paraffin-embedded sections from 6-month-old $p48^{Cre};Kras^{G12D};Rip3^{+/+}$ (KC;$Rip3^{+/+}$) and $p48^{Cre};Kras^{G12D};Rip3^{-/-}$ mice were tested for expression of CXCL1 by IHC ($n=3$ per group). **g**, Homogenized pancreas tissue from 6-month-old wild-type (WT) ($n=6$), $p48^{Cre};Kras^{G12D};Rip3^{+/+}$ ($n=4$), and $p48^{Cre};Kras^{G12D};Rip3^{-/-}$ ($n=4$) mice was tested for CXCL1 by ELISA. **h**, $Kras^{G12D};Rip3^{+/+}$ PDEC ($n=3$) and $Kras^{G12D};Rip3^{-/-}$ PDEC ($n=4$) were cultured *in vitro* for 24 h and the supernatant was tested for CXCL1 by ELISA. PDEC, pancreatic ductal epithelial cells. **i**, Correlation between high and low tertiles of *RIP3* expression and *CXCL1* expression in human PDA tissues tested using the UCSC RNA-seq database. Each point represents data from one patient. **j**, Wild-type or $Rip3^{-/-}$ mice were orthotopically implanted with KPC-derived tumour cells and treated with a single dose of gemcitabine or PBS 3 weeks later. Tumours were removed 12 h after treatment and homogenized, and the fold-difference in CXCL1 expression between gemcitabine- and PBS-treated tumours from wild-type and $Rip3^{-/-}$ mice was determined by ELISA ($n=5$). **k**, KPC-derived tumour cells were treated in triplicate *in vitro* with gemcitabine (10 μM or 50 μM), Nec-1s, or a RIP3 inhibitor (RIP3-Inh.), alone or in combination. CXCL1 levels were tested after 24 h by ELISA. Experiments were repeated at least twice. Graphs show mean ± s.e.m. n.s., not significant; $*P<0.05$, $**P<0.01$, $***P<0.001$, $****P<0.0001$ (unpaired *t*-test).

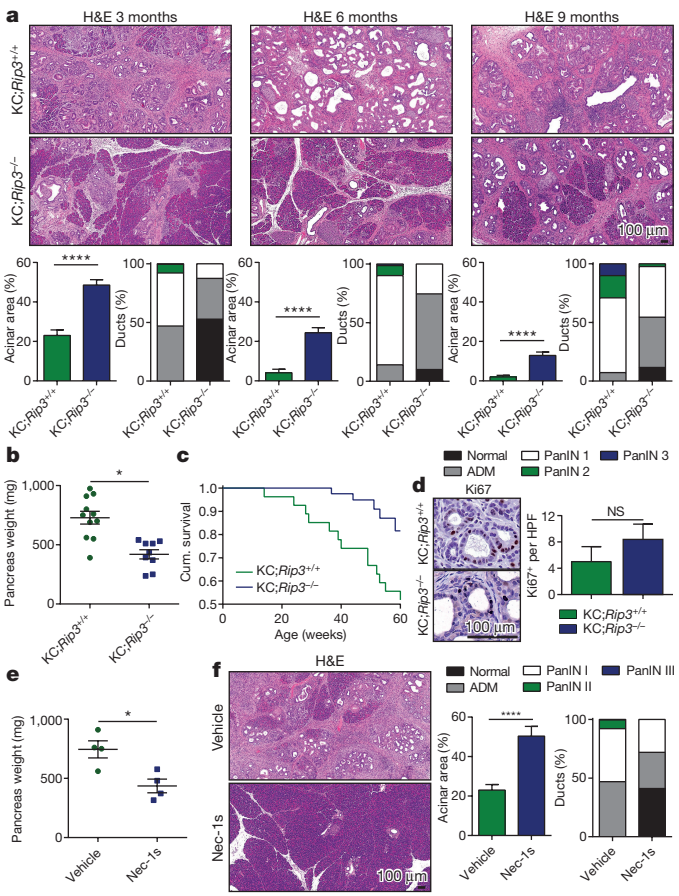


Figure 3 | Deletion of RIP3 or blockade of RIP1 protects against pancreatic oncogenesis. **a**, $p48^{Cre};Kras^{G12D};Rip3^{+/+}$ ($n=11$) and $p48^{Cre};Kras^{G12D};Rip3^{-/-}$ ($n=9$) mice were killed at 3, 6, or 9 months of age. Representative haematoxylin and eosin (H&E)-stained sections, the percentage of pancreatic area occupied by intact acinar structures, and the fractions of ductal structures exhibiting normal morphology, acino-ductal metaplasia (ADM), or graded PanIN I–III lesions are shown. **b**, Pancreas weights were compared in 3-month-old $p48^{Cre};Kras^{G12D};Rip3^{+/+}$ ($n=11$) and $p48^{Cre};Kras^{G12D};Rip3^{-/-}$ ($n=9$) mice. **(c)** Kaplan–Meier survival analysis was performed for $p48^{Cre};Kras^{G12D};Rip3^{+/+}$ ($n=29$) and $p48^{Cre};Kras^{G12D};Rip3^{-/-}$ ($n=41$) mice ($P<0.005$). **d**, Pancreases from 3-month-old $p48^{Cre};Kras^{G12D};Rip3^{+/+}$ and $p48^{Cre};Kras^{G12D};Rip3^{-/-}$ mice were assayed for expression of Ki67. **e**, **f**, Six week-old KC mice were serially treated with the RIP1 inhibitor Nec-1s or vehicle for 8 weeks before being killed ($n=4$ per group). Pancreas weights (**e**) and representative H&E-stained sections are shown and ductal morphology was quantified (**f**). Nec-1s experiments were repeated three times with similar results. Graphs show mean ± s.e.m. $*P<0.05$; $****P<0.0001$ (unpaired *t*-test).

pro-tumorigenic immune suppression associated with RIP3 signalling by mobilizing myeloid cells[14,15]. To test this hypothesis, we challenged wild-type mice with orthotopic PDA while blocking CXCL1. CXCL1 blockade protected against tumorigenesis induced by either orthotopic $Kras^{G12D}$ PDEC (data not shown) or KPC cells (Extended Data Fig. 4c). However, anti-CXCL1 treatment did not further enhance tumour protection in $Rip3^{-/-}$ mice (Extended Data Fig. 4d). Moreover, like RIP3 deletion, CXCL1 blockade reduced MDSC and TAM accumulation (Extended Data Fig. 4e, f). Tumour-infiltrating T cells were also more activated in mice in which CXCL1 was blocked than in control mice, as shown by higher CD44 and TNFα expression (Extended Data Fig. 4g, h). However, CXCL1 inhibition was not significantly associated with increased infiltration of peri-tumoral T cells (Extended Data Fig. 4i); nor did it diminish T_{reg} cell accumulation or IL-10 expression (data not shown), unlike RIP3 deletion. Together, these data suggest that CXCL1 overexpression alone may

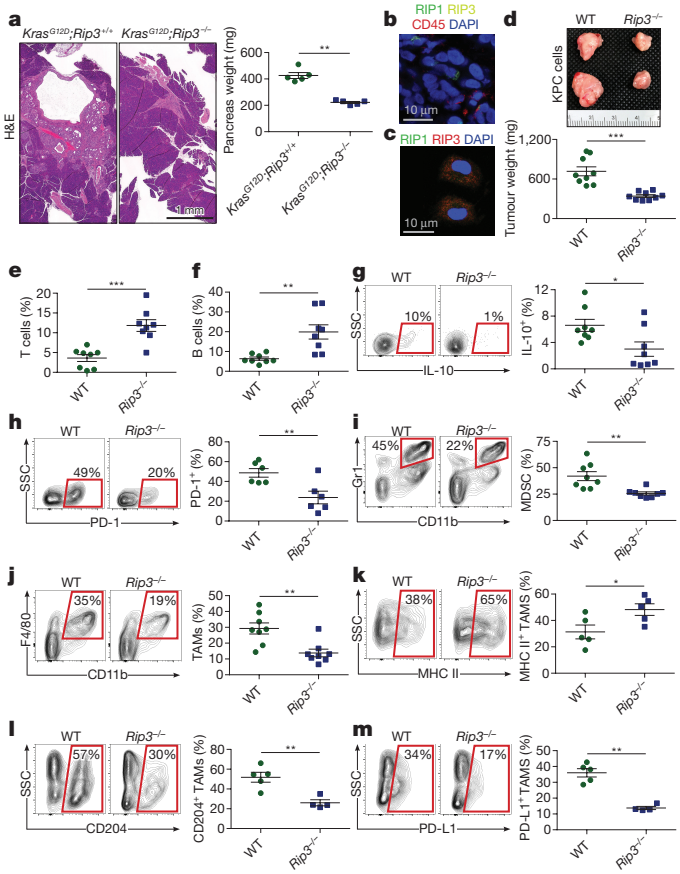


Figure 4 | RIP3 deletion in the epithelial or extra-epithelial compartment protects against PDA and enhances immunogenicity. **a**, Wild-type (WT) mice orthotopically implanted with $Kras^{G12D}$;$Rip3^{+/+}$ or $Kras^{G12D}$;$Rip3^{-/-}$ PDEC were killed at 6 weeks. Representative H&E-stained sections and pancreas weights are shown ($n = 5$ per group). **b**, Frozen sections of orthotopic KPC tumours were co-stained for RIP1, RIP3, and CD45 and imaged by immune-fluorescence microscopy. **c**, $CD45^+$ leukocyte suspensions harvested from orthotopic KPC tumours were co-stained for RIP1 and RIP3 and imaged by immune-fluorescence microscopy. **d**, WT and $RIP3^{-/-}$ mice bearing orthotopic KPC tumours were killed at 3 weeks. Representative gross pictures and tumour weights are shown ($n = 9$ per group). **e**–**h**, The fractions of peri-tumoral $CD3^+$ T cells (**e**) and $CD19^+$ B cells (**f**) and the T cell expression of IL-10 (**g**) and PD-1 (**h**) from orthotopic KPC tumours. SSC, side scatter. **i**, **j**, The fraction of peri-tumoral $Gr1^+CD11b^+$ MDSC (**i**) and $Gr1^-CD11b^+F4/80^+$ TAMs (**j**). **k**–**m**, Expression of MHC II (**k**), CD204 (**l**), and PD-L1 (**m**) on TAMs. Flow cytometry data were reproduced in three separate experiments. Graphs show mean ± s.e.m. $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$ (unpaired *t*-test).

not account for the entire immunosuppressive phenotype associated with intact necroptosis signalling in PDA.

We postulated that necroptotic tumour cells release soluble factors that induce peri-tumoral immune suppression. Mincle, a C-type lectin receptor (CLR) that is crucial for immunity to mycobacteria, can promote sterile inflammation *in vitro* by ligating SAP130, a nuclear protein released from dying cells[16,17]. We discovered high cytoplasmic SAP130 expression in human PDA (Extended Data Fig. 5a). *SAP130* expression was also upregulated by gemcitabine treatment in human PDA cell lines (Extended Data Fig. 5b). Furthermore, SAP130 was highly expressed in $p48^{Cre}$;$Kras^{G12D}$;$Rip3^{+/+}$ pancreases, whereas its expression was reduced in $p48^{Cre}$;$Kras^{G12D}$;$Rip3^{-/-}$ pancreases (Extended Data Fig. 5c). Similarly, gemcitabine-induced upregulation of *Sap130* was reduced by Nec-1s in mouse PDA cells (Extended Data Fig. 5d). SAP130 was expressed in both epithelial and inflammatory cells of PDA (Extended Data Fig. 5e, f). Moreover, confocal microscopy suggested that SAP130 co-localized with RIP1/RIP3 in human (Extended Data Fig. 5g) and mouse (data not shown) PDA cells. *SAP130* expression also correlated with *RIP1–RIP3* co-expression in a human RNA-seq database (Extended Data Fig. 5h). Notably, there was a trend towards an association between high *SAP130* expression and low survival in humans with PDA (Extended Data Fig. 5i).

We postulated that ligation of Mincle by SAP130 drives necrosome-induced accelerated oncogenesis. Consistent with this idea, immunoprecipitation experiments indicated that Mincle is associated with SAP130 in PDA (Extended Data Fig. 5j). Mincle was expressed in inflammatory cells in the human PDA TME, but not in transformed ductal cells or in normal pancreas cells (Extended Data Fig. 6a). Overall, around 10% of tumour-infiltrating leukocytes expressed Mincle in human PDA compared with <1% of peripheral blood mononuclear cells (PBMC; Extended Data Fig. 6b). Subset analysis revealed that Mincle expression was higher in human $CD14^+CD15^+$ tumour-infiltrating monocytes than in PBMC (Extended Data Fig. 6c). Similarly, in KC mice 10–15% of pancreatic leukocytes expressed Mincle compared with ~3% expression in the spleen and <1% expression in pancreatic parenchymal cells (Extended Data Fig. 6d). Immunofluorescence microscopy confirmed that Mincle was expressed in enriched PDA-infiltrating leukocytes (Extended Data Fig. 6e). Subset analysis indicated that Mincle was more highly expressed in PDA-infiltrating MDSC, dendritic cells, and macrophages than in these cellular subsets in the spleen (Extended Data Fig. 6f). Western blotting showed RIP3-dependent elevated expression of Mincle-related signalling intermediates in PDA (Extended Data Fig. 6g). Accordingly, there were few phosphorylated Syk^+ leukocytes in $p48^{Cre}$;$Kras^{G12D}$;$Rip3^{-/-}$ and $p48^{Cre}$;$Kras^{G12D}$;$Mincle^{-/-}$ (*Mincle* is also known as *Clec4e*) pancreases compared to $p48^{Cre}$;$Kras^{G12D}$;$Rip3^{+/+}$;$Mincle^{+/+}$ pancreases (Extended Data Fig. 6h). However, deletion of Mincle in PDA did not reduce CXCL1 expression (Extended Data Fig. 6i), and CXCL1 blockade did not alter the expression of Mincle-associated signalling intermediates (data not shown).

To determine whether Mincle signalling accelerates oncogenesis, we treated 6-week-old KC mice thrice weekly with the Mincle ligand TDB (trehalose-6,6-dibehenate), which we confirmed induced phosphorylation of Syk *in vivo* (Extended Data Fig. 7a). Mincle ligation accelerated tumorigenesis, which resulted in higher grade PanIN lesions, extensive fibrosis, and scattered foci of invasion when compared with control mice (Extended Data Fig. 7b). TDB also accelerated the growth of orthotopically implanted KPC-derived tumours in wild-type (not shown) and $Rip3^{-/-}$ mice (Extended Data Fig. 7c), suggesting that the pro-tumorigenic effects of Mincle activation in PDA are either independent of or downstream of RIP3 signalling. Moreover, the inflammatory TME in TDB-treated $Rip3^{-/-}$ pancreases recapitulated the immune-suppressive milieu associated with intact necroptosis signalling. Specifically, TDB-treated pancreases trended to contain a lower fraction of tumour-infiltrating T cells (Extended Data Fig. 7d) and exhibited increased recruitment of both MDSC (Extended Data Fig. 7e) and M2-like TAMs, which expressed high PD-L1, compared with control pancreases (Extended Data Fig. 7f–i). Similarly, direct inoculation of orthotopic PDA tumours with recombinant SAP130 accelerated PDA growth in wild-type and $Rip3^{-/-}$ mice but not in $Mincle^{-/-}$ mice (Extended Data Fig. 7j) and recruited an immune-suppressive infiltrate in WT mice (Extended Data Fig. 7k, l).

To determine whether Mincle signalling is required for PDA progression, we crossed $Mincle^{-/-}$ and KC mice and analysed pancreases at 3 month intervals. Mincle deletion slowed the rate of oncogenesis based on histological analysis, pancreas weight, and animal survival (Extended Data Fig. 8a–c). Similarly, orthotopic KPC-derived tumour implantation in the pancreases of $Mincle^{-/-}$ mice resulted in smaller tumours and prolonged survival compared with implantation in wild-type mice (Extended Data Fig. 8d, e). However, the survival

benefit was not as pronounced in *Mincle*$^{-/-}$ mice as in *Rip3*$^{-/-}$ mice. Moreover, experiments involving orthotopic implantation of *Kras*^{G12D} PDEC suggested that combined blockade of both CXCL1 and Mincle had additive protective effects over RIP3 blockade alone, whereas combined blockade of RIP3 and Mincle or RIP3 and CXCL1 did not (Extended Data Fig. 8f).

To determine whether Mincle deletion mimics the immunogenic reprogramming of the TME that is associated with RIP3 deletion, we assayed the pancreatic infiltrate in *p48*^{Cre};*Kras*^{G12D};*Mincle*$^{-/-}$ mice. Immunohistochemical (IHC) analysis of *p48*^{Cre};*Kras*^{G12D};*Mincle*$^{-/-}$ pancreases showed diminished TAM infiltration but increased T cell recruitment (Extended Data Fig. 9a). Flow cytometry confirmed that Mincle deletion was associated with increased immunogenic T cell infiltration (Extended Data Fig. 9b–d), diminished MDSC infiltration (Extended Data Fig. 9e), a trend towards a reduction in the number of dendritic cells (Extended Data Fig. 9f), and a decreased fraction of TAMs (Extended Data Fig. 9g) with M1-like polarization (Extended Data Fig. 9h, i) and reduced PD-L1 expression (data not shown). These changes recapitulate the immunogenic reprogramming of the TME that follows RIP3 deletion.

To investigate whether protection against oncogenesis in the absence of RIP3 or Mincle signalling is T cell dependent, we depleted T cells and implanted orthotopic KPC tumours in wild-type, *Rip3*$^{-/-}$, and *Mincle*$^{-/-}$ mice. T cell depletion did not influence PDA growth in wild-type mice. However, protection against tumour growth was abrogated by T cell depletion in *Rip3*$^{-/-}$ and *Mincle*$^{-/-}$ mice (Extended Data Fig. 10a). Conversely, depletion of macrophages in wild-type mice led to T cell activation and protection against tumour growth; however, macrophage depletion did not enhance T cell activation or induce protection against oncogenesis in *Rip3*$^{-/-}$ or *Mincle*$^{-/-}$ mice (Extended Data Fig. 10b, c). These data suggest that, in wild-type mice bearing orthotopic tumours, TAMs promote PDA progression and T cells are not tumour-protective; conversely, in the absence of RIP3 or Mincle signalling, macrophages surrender their tumour-promoting effects and T cells are reprogrammed into indispensable mediators of anti-tumour immunity. Collectively, our findings indicate that necroptosis-induced CXCL1 and Mincle signalling promotes myeloid cell-induced adaptive immune suppression in PDA. Each of these networks represents a novel target for experimental therapeutic agents (Extended Data Fig. 10d).

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 4 May 2015; accepted 5 February 2016.
Published online 6 April 2016.

1. Johnstone, R. W., Ruefli, A. A. & Lowe, S. W. Apoptosis: a link between cancer genetics and chemotherapy. *Cell* **108,** 153–164 (2002).
2. Fernald, K. & Kurokawa, M. Evading apoptosis in cancer. *Trends Cell Biol.* **23,** 620–633 (2013).
3. Lowe, S. W. & Lin, A. W. Apoptosis in cancer. *Carcinogenesis* **21,** 485–495 (2000).
4. Holler, N. *et al.* Fas triggers an alternative, caspase-8-independent cell death pathway using the kinase RIP as effector molecule. *Nature Immunol.* **1,** 489–495 (2000).
5. Vanden Berghe, T., Linkermann, A., Jouan-Lanhouet, S., Walczak, H. & Vandenabeele, P. Regulated necrosis: the expanding network of non-apoptotic cell death pathways. *Nature Rev. Mol. Cell Biol.* **15,** 135–147 (2014).
6. Vandenabeele, P., Galluzzi, L., Vanden Berghe, T. & Kroemer, G. Molecular mechanisms of necroptosis: an ordered cellular explosion. *Nature Rev. Mol. Cell Biol.* **11,** 700–714 (2010).
7. Nagy, A. *et al.* Copy number of cancer genes predict tumor grade and survival of pancreatic cancer patients. *Anticancer Res.* **21,** 1321–1325 (2001).
8. Plath, T. *et al.* Overexpression of pRB in human pancreatic carcinoma cells: function in chemotherapy-induced apoptosis. *J. Natl. Cancer Inst.* **94,** 129–142 (2002).
9. Rosty, C. *et al.* p16 Inactivation in pancreatic intraepithelial neoplasias (PanINs) arising in patients with chronic pancreatitis. *Am. J. Surg. Pathol.* **27,** 1495–1501 (2003).
10. Takahashi, H. *et al.* Simultaneous knock-down of Bcl-xL and Mcl-1 induces apoptosis through Bax activation in pancreatic cancer cells. *Biochim. Biophys. Acta* 1833, 2980–2987 (2013).
11. Ochi, A. *et al.* Toll-like receptor 7 regulates pancreatic carcinogenesis in mice and humans. *J. Clin. Invest.* **122,** 4118–4129 (2012).
12. He, S. *et al.* Receptor interacting protein kinase-3 determines cellular necrotic response to TNF-α. *Cell* **137,** 1100–1111 (2009).
13. Yatim, N. *et al.* RIPK1 and NF-κB signaling in dying cells determines cross-priming of CD8$^+$ T cells. *Science* **350,** 328–334 (2015).
14. Connolly, M. K. *et al.* Distinct populations of metastases-enabling myeloid cells expand in the liver of mice harboring invasive and preinvasive intra-abdominal tumor. *J. Leukoc. Biol.* **87,** 713–725 (2010).
15. Acharyya, S. *et al.* A CXCL1 paracrine network links cancer chemoresistance and metastasis. *Cell* **150,** 165–178 (2012).
16. Yamasaki, S. *et al.* Mincle is an ITAM-coupled activating receptor that senses damaged cells. *Nature Immunol.* **9,** 1179–1188 (2008).
17. Wells, C. A. *et al.* The macrophage-inducible C-type lectin, mincle, is an essential component of the innate immune response to *Candida albicans*. *J. Immunol.* **180,** 7404–7413 (2008).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by grants from the German Research Foundation (L.S.), the National Pancreas Foundation (C.P.Z.), the Pancreatic Cancer Action Network (G.M.), the Lustgarten Foundation (G.M.), and National Institute of Health Awards CA155649 (G.M.), CA168611 (G.M.), and CA193111 (G.M., A.T.-H.). We thank the New York University Langone Medical Center (NYU LMC) Histopathology Core Facility, the NYU LMC Flow Cytometry Core Facility, the NYU LMC Microscopy Core Facility, and the NYU LMC BioRepository Center, each supported in part by the Cancer Center Support Grant P30CA016087 and by grant UL1 TR000038 from the National Center for the Advancement of Translational Science (NCATS).

Author Contributions L.S. carried out *in vivo* experiments, flow cytometry, analysis and interpretation, manuscript preparation, and statistical analysis; G.W. carried out *in vivo* experiments, flow cytometry, analysis and interpretation, manuscript preparation, and statistical analysis; S.T. carried out *in vivo* experiments and IHC; N.N.G.L. performed western blotting; S.A. carried out IHC; D.A. performed flow cytometry; A.A. performed tissue culture and cell line generation; R.B. provided technical assistance and critical review; D.D. performed flow cytometry and provided critical review; S.H.G. carried out mouse breeding and provided critical review; A.T.-H. provided technical assistance and critical review; M.P. performed western blotting and flow cytometry and provided critical review; A.O. carried out immunoprecipitation; C.P.Z. provided technical advice and performed PCR and flow cytometry; M.P. performed western blotting; M.R. carried out genotyping; D.T. performed animal breeding and *in vivo* tumour experiments; C.H. carried out histological analysis; M.H. performed FACS and data analysis; V.R.M. performed FACS and data analysis; D.E. created cell lines and performed *in vivo* experiments; G.M. conceived, designed, supervised, analysed and interpreted the study and provided critical review.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to G.M. (george.miller@nyumc.org).

METHODS

Animals and *in vivo* models. C57BL/6 (H-2Kb) mice were purchased from Jackson Labs (Bar Harbour, ME). *Mincle*$^{-/-}$ mice were obtained from the MMRRC (San Diego, CA)[17]. *Rip3*$^{-/-}$ mice were obtained from Genentech (San Francisco, CA)[18]. KC (gift from D. Bar-Sagi, New York University) and KPC (gift from M. Philips, New York University) mice develop pancreatic neoplasia endogenously by expressing mutant *Kras* alone or mutant *Kras* and *p53*, respectively, in the progenitor cells of the pancreas[19,20]. Both male and female mice were used but animals were sex- and age-matched in each experiment. Randomization was not performed. There were no specific inclusion or exclusion criteria. Sample sizes for experiments were determined without formal power calculations. Survival data for control KC mice have been previously reported[21]. For orthotopic pancreatic tumour challenge, mice were given intra-pancreatic injections of either *Kras*^{G12D} PDEC or FC1242 tumour cells derived from KPC mice. *Kras*^{G12D} PDEC and FC1242 cells were generated as previously described[21,22]. In preparation for intra-pancreatic injection, cells were suspended in PBS with 50% Matrigel (BD Biosciences, Franklin Lakes, NJ) at 1×10^6 cells ml^{-1} and 1×10^5 cells were injected into the body of the pancreas via laparotomy. Age-matched mice were used between 8 and 10 weeks of age for orthotopic tumour experiments. Mice were killed by cervical dislocation 3–6 weeks later and tumour volume recorded. To study the effects of Mincle ligation, mice were injected intraperitoneally (i.p.) with TDB (4 mg kg^{-1}; Invivogen, San Diego, CA) thrice weekly for 8 weeks in the endogenous tumour models and for 3 weeks in the orthotopic tumour models. In other experiments, orthotopic tumours were serially treated with direct inoculation of recombinant SAP130 (22 μg; MyBioSource, San Diego, CA) at one-week intervals via mini-laparotomy. In select experiments, cohorts of mice were treated daily with the RIP1 inhibitor Nec-1s (2 mg kg^{-1}, i.p.; BioVision, Milpitas, CA) or a neutralizing anti-CXCL1 monoclonal antibody (mAb) (4 mg kg^{-1}, i.v.; R&D Systems). Gemcitabine (100 mg kg^{-1}, i.p.; Hospira, Lake Forest, IL) was administered *in vivo* to KPC mice three times at 72-h intervals unless otherwise specified. T cells (T24/31) and macrophages (F4/80, both BioXcell) were depleted with neutralizing mAbs as previously described[23]. In some experiments, mice were subcutaneously administered FC1242 cells (1×10^6) or B16 melanoma cells (1×10^6; gift from R. DeMatteo, Memorial Sloan-Kettering Cancer Center) and killed 18 days later. Investigators were not blinded to group allocation but were blinded when assessing outcome. All animal procedures were approved by the New York University School of Medicine IACUC. The maximum tumour size permitted was 3 cm^3 and this was not exceeded.

Cell lines and *in vitro* experiments. The human PDA cell lines AsPC-1, PANC1, and MIA PaCa-2 (gifts from D. Bar-Sagi, originally obtained from ATCC) were maintained in complete RPMI (RPMI 1640 with 10% heat-inactivated FBS, 2 mM L-glutamine, 1% penicillin/streptomycin). Cell lines were not authenticated. Cells were free of mycoplasma. In selected experiments, cells were treated with gemcitabine (10–50 μM), Nec-1s (50 μM), a RIP3 inhibitor (GSK872; 6 μM), or a MLKL inhibitor (necrosulphonamide, 1 μM, both EMD Millipore, Billerica, MA). Cell viability was determined by PI staining. Cellular proliferation was assessed using the XTT II assay according to the manufacturer's protocol (Roche, Pleasanton, CA) and expressed as per cent proliferation compared to control. Inflammatory mediators in cell culture supernatant were measured using the Milliplex Immunoassay (Millipore, Billerica, MA). CXCL1 was additionally measured using Flexbeads (BD Biosciences) and ELISA (R&D Systems).

Cellular harvest and flow cytometry. Human and mouse single cell suspensions for flow cytometry were prepared as described previously with slight modifications[24]. Briefly, pancreases were placed in cold RPMI 1640 with 1 mg ml^{-1} collagenase IV (Worthington Biochemical, Lakewood, NJ) and 2 U ml^{-1} DNase I (Promega, Madison, WI) and minced with scissors to sub-millimetre pieces. Tissues were then incubated at 37 °C for 30 min with gentle shaking every 5 min. Specimens were passed through a 100-μm mesh, and centrifuged at 350*g* for 5 min. The cell pellet was resuspended in cold PBS with 1% FBS. After blocking FcγRIII/II with an anti-CD16/CD32 mAb (eBioscience, San Diego, CA), cell labelling was performed by incubating 10^6 cells with 1 μg fluorescently conjugated mAbs directed against mouse CD44 (IM7), CD206 (C068C2), PD-L1 (10F.9G2), PD-1 (29F.1A12), CD3 (17A2), CD4 (RM4-5), CD8 (53-6.7), CD45 (30-F11), CD11b (M1/70), Gr1 (RB6-8C5), CD11c (N418), MHC II (M5/114.15.2), IL-10 (JES5-16E3), IFN-γ (XMG1.2), TNFα (MP6-XT22), F4/80 (BM8), CD19 (6D5; all Biolegend, San Diego, CA), p-Syk (moch1ct, eBioscience), and CD204 (2F8; Acris Antibodies, San Diego, CA). mAbs directed against Mincle (4A9, MBL International Corporation, Woburn, MA) were conjugated to FITC using the FITC Conjugation Kit (Abcam, Cambridge, MA). Human pancreas and PBMC were stained with mAbs directed against CD45 (HI30), CD14 (HCD14), CD15 (W6D3), CD19 (H1B19), CD11b (M1/70), CD11c (3.9), MHC II (L243; all Biolegend) and Mincle (AT16E3; Acris Antibodies). Intracellular cytokine staining was performed using the Fixation/Permeabilization Solution Kit (BD Biosciences). Flow cytometry was carried out on the LSR-II flow cytometer (BD Biosciences). Data were analysed using FlowJo v.7.6.5 (Treestar, Ashland, OR).

Western blotting and immunoprecipitation. To extract proteins from tissues, 15–30 mg of tissue was homogenized in 150–300 μl (that is, 10 times the weight) ice-cold RIPA buffer. Total protein was quantified using the BioRad DC Protein Assay according to the manufacturer's instructions (BioRad, Hercules, CA). Western blotting was performed as described previously with minor modifications[24]. Briefly, 10% Bis-Tris polyacrylamide gels (NuPage, Invitrogen) were equiloaded with 10–30 μg protein, electrophoresed at 200 V, and electro-transferred to PVDF membranes. After blocking with 5% BSA, membranes were probed with primary antibodies to β-actin (8H10D10), FADD (polyclonal), RIP1 (D94C12), caspase-8 (D35G2), PLC-γ (polyclonal), p-PLC-γ (polyclonal), Bcl-xL (54H6; all Cell Signaling, Beverly, MA), RIP3 (polyclonal; Abgent, San Diego, CA), c-Myc (9E10), CDK4 (C-22), CARD9 (polyclonal), Syk (polyclonal), p-Syk (polyclonal), Rb (C-15), SAP130 (H-300), Mincle (H-46; all Santa Cruz Biotechnologies, Dallas, TX), and MLKL (polyclonal; Abcam). Blots were developed by ECL (Thermo Scientific, Asheville, NC). For immunoprecipitation experiments, RIP1 or SAP130 was precipitated with protein G-agarose from cells. Immunoprecipitates were re-suspended and heated in loading buffer under reduced conditions, and resolved by 10% SDS–PAGE before transfer to PVDF membranes. The presence of co-immunoprecipitated RIP3 or Mincle, respectively, was determined by western blotting.

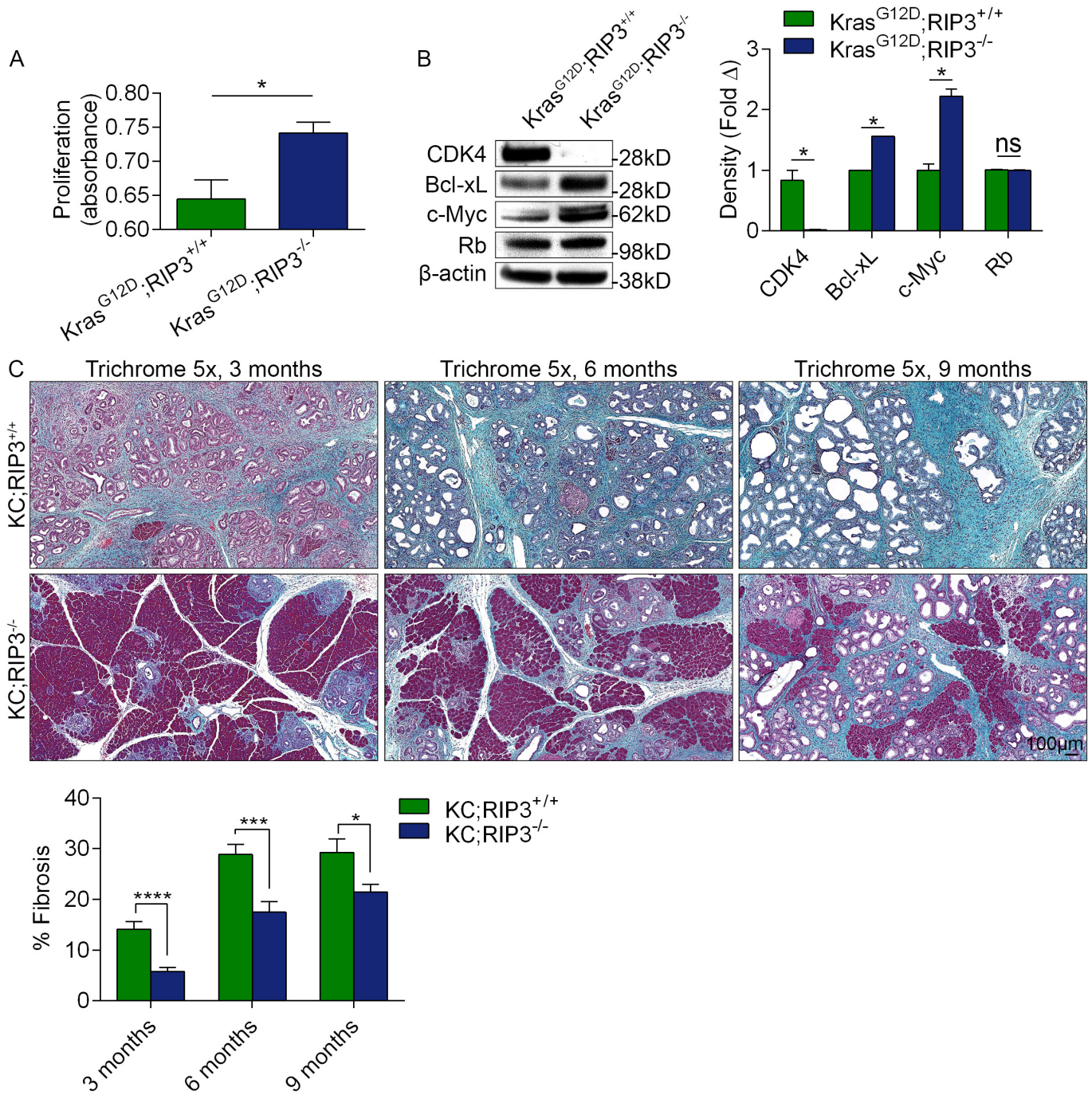
Histology, immunohistochemistry, and microscopy. For histological analysis, pancreatic specimens were fixed with 10% buffered formalin, dehydrated in ethanol, embedded with paraffin, and stained with H&E or Gomori's trichrome. The fraction of preserved acinar area was calculated as previously described[24]. Pancreatic ductal dysplasia was graded according to established criteria[25]. Immunohistochemistry in mouse tissues was performed using antibodies directed against F4/80 (CI:A3-1; Abcam), CD3 (polyclonal; Abcam), Arg1 (EPR6671(B); Abcam), SAP130 (polyclonal; Abcam), Mincle (AT16E3; Abcam), p-Syk (polyclonal; Abcam), Ki67 (polyclonal; Abcam), and CXCL1 (polyclonal, Abcam). For analysis of human tissue, de-identified paraffin-embedded human PDA specimens and samples of surrounding non-tumorous tissue from ten consecutive patients who underwent surgical resection of PDA at NYU Medical Center were probed with antibodies directed against RIP1 (D94C12; Cell Signaling), RIP3 (Q9Y572; Abgent), Mincle (AT16E3; Abcam), CXCL1 (polyclonal; Abcam), and SAP130 (polyclonal; Abcam). All human tissues were collected using an IRB approved protocol and donors of de-identified specimens gave informed consent. Sample sizes for human experiments were not determined based on formal power calculations. Quantifications were performed by assessing ten high-power fields (HPF; 40×) per slide in a blinded manner. Immunofluorescent staining of frozen mouse tissues or cells was performed using antibodies against Mincle (AT16E3, Acris Antibodies), RIP1 (polyclonal, Bioss), RIP3 (polyclonal, Bioss), CD45 (30-F11; BD Biosciences), CK19 (clone 13, Abnova), CXCL1 (polyclonal; Abcam), CXCR2 (SA045E1; BioLegend), SAP130 (polyclonal; Abcam), and DAPI (Vector Labs, Burlingame, CA). Immunofluorescent images were acquired using a Zeiss LSM700 confocal microscope with ZEN 2010 software (Carl Zeiss, Thornwood, New York).

PCR. RNA was extracted using the RNeasy Mini kit (Qiagen, Germantown, MD) according to the manufacturer's instructions. RNA was converted to cDNA using the RT2 First Strand Kit (Qiagen). qPCR was performed using RT2 SYBR Green qPCR mastermix (Qiagen) on a Stratagene MX3005P (Stratagene, La Jolla, CA) according to the manufacturers' protocols. Primers used for human and mouse samples (*RIP1*, *RIP3*, *CASP8*, *FADD*, *CXCL1*, and *SAP130*) were purchased from Qiagen. Expression levels were normalized to β-actin (*ACTβ*) and expressed as fold change compared to control.

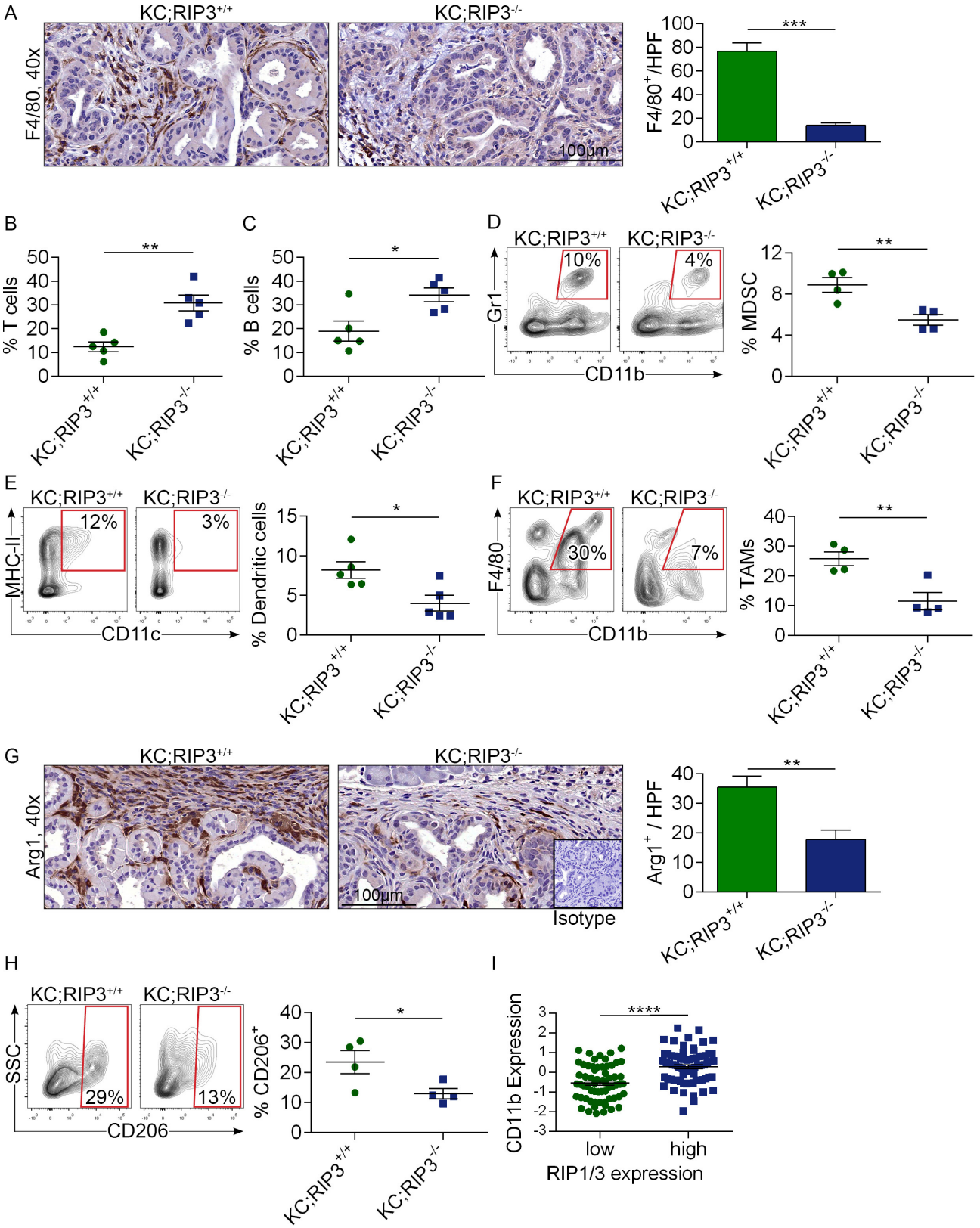
Human database and statistical analysis. Human RNA-seq data and clinical correlations were performed using the UCSC Cancer Genomics Browser (https://genome-cancer.ucsc.edu/)[26]. Data are presented as mean ± s.e.m. Survival was measured using the Kaplan–Meier method. Statistical significance was determined by Student's *t*-test and the log-rank test using GraphPad Prism 6 (GraphPad Software, La Jolla, CA). $P < 0.05$ was considered significant.

18. Newton, K., Sun, X. & Dixit, V. M. Kinase RIP3 is dispensable for normal NF-κBs, signaling by the B-cell and T-cell receptors, tumor necrosis factor receptor 1, and Toll-like receptors 2 and 4. *Mol. Cell. Biol.* **24,** 1464–1469 (2004).
19. Hingorani, S. R. *et al.* Preinvasive and invasive ductal pancreatic cancer and its early detection in the mouse. *Cancer Cell* **4,** 437–450 (2003).
20. Hingorani, S. R. *et al.* Trp53^{R172H} and Kras^{G12D} cooperate to promote chromosomal instability and widely metastatic pancreatic ductal adenocarcinoma in mice. *Cancer Cell* **7,** 469–483 (2005).

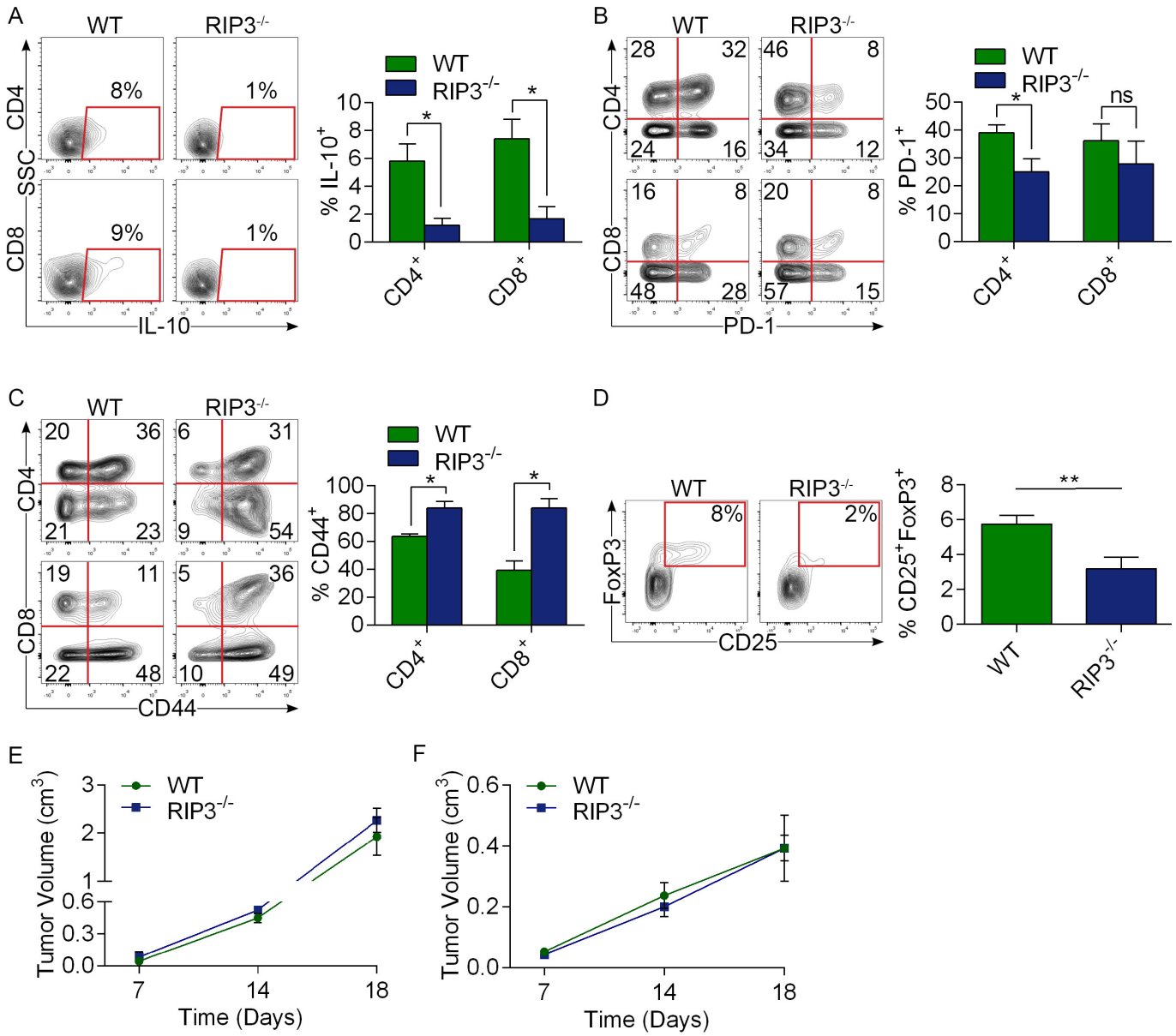
21. Zambirinis, C. P. *et al.* TLR9 ligation in pancreatic stellate cells promotes tumorigenesis. *J. Exp. Med.* **212,** 2077–2094 (2015).
22. Pylayeva-Gupta, Y., Lee, K. E., Hajdu, C. H., Miller, G. & Bar-Sagi, D. Oncogenic Kras-induced GM-CSF production promotes the development of pancreatic neoplasia. *Cancer Cell* **21,** 836–847 (2012).
23. Bedrosian, A. S. *et al.* Dendritic cells promote pancreatic viability in mice with acute pancreatitis. *Gastroenterology* **141,** 1915–1926 (2011).
24. Ochi, A. *et al.* MyD88 inhibition amplifies dendritic cell capacity to promote pancreatic carcinogenesis via Th2 cells. *J. Exp. Med.* **209,** 1671–1687 (2012).
25. Hruban, R. H. *et al.* Pancreatic intraepithelial neoplasia: a new nomenclature and classification system for pancreatic duct lesions. *Am. J. Surg. Pathol.* **25,** 579–586 (2001).
26. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).



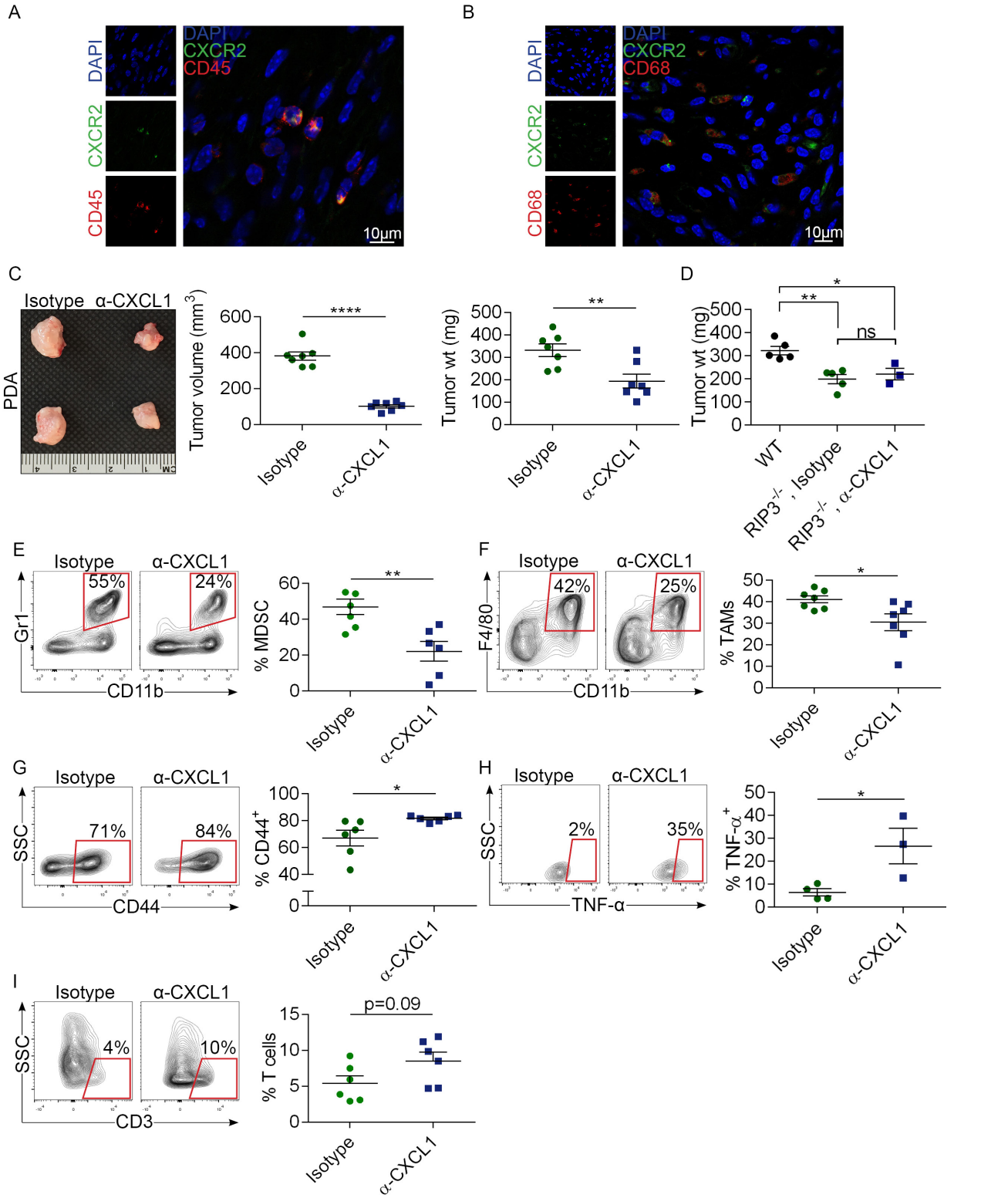
Extended Data Figure 1 | RIP3 deletion in PDA induces an aggressive tumour phenotype *in vitro* but mitigates oncogenesis *in vivo*. **a**, $Kras^{G12D};Rip3^{+/+}$ and $Kras^{G12D};Rip3^{-/-}$ PDEC were cultured at equal densities and tested for proliferation after 24 h using the XTT assay ($n = 6$ per group). **b**, Lysate was harvested from $Kras^{G12D};Rip3^{+/+}$ and $Kras^{G12D};Rip3^{-/-}$ PDEC and tested for expression of selected tumour suppressor and oncogenic genes. Representative data and density plots from biological duplicates are shown. Experiments were reproduced three times. **c**, $p48^{Cre};Kras^{G12D};Rip3^{+/+}$ ($n = 11$) and $p48^{Cre};Kras^{G12D};Rip3^{-/-}$ ($n = 9$) mice were killed at 3, 6, or 9 months of age. Representative trichrome-stained sections are shown and the fraction of fibrotic pancreatic area was calculated for each cohort. Graphs show mean ± s.e.m. ns, not significant; $*P < 0.05$, $***P < 0.001$, $****P < 0.0001$ (unpaired *t*-test). For gel source data, see Supplementary Fig. 1.



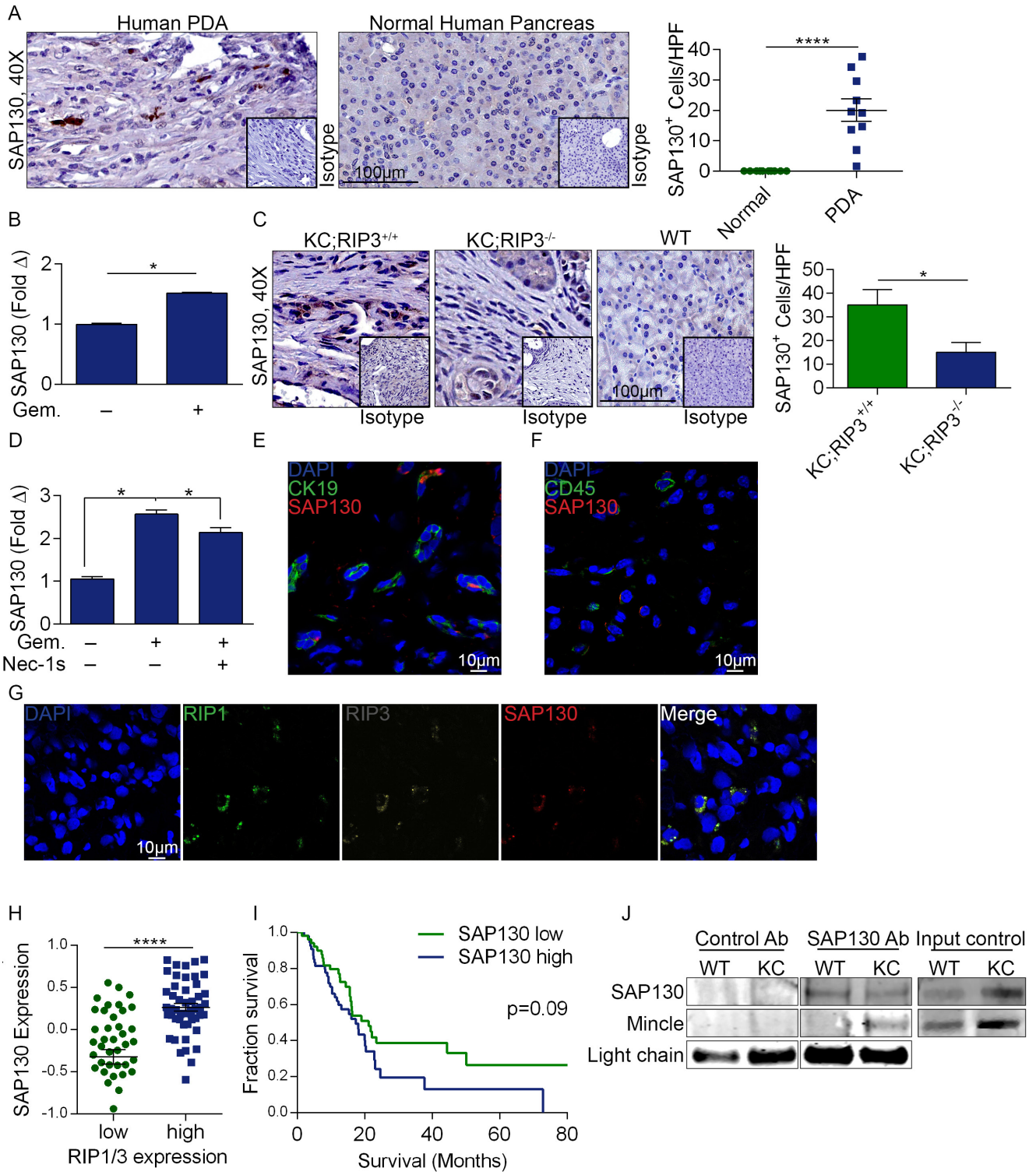
Extended Data Figure 2 | RIP3 deletion induces immunogenic reprogramming of the pancreatic TME. a, $p48^{Cre}$;$Kras^{G12D}$;$Rip3^{+/+}$ and $p48^{Cre}$;$Kras^{G12D}$;$Rip3^{-/-}$ mice were killed at 3 months of age. Paraffin-embedded sections were stained using a mAb directed against F4/80. Representative images and quantitative data are shown ($n = 5$ per group). **b–f**, The fraction of peri-tumoral $CD3^+$ T cells (**b**), $CD19^+$ B cells (**c**), $Gr1^+CD11b^+$ MDSC (**d**), $F4/80^-CD11c^+MHCII^+$ dendritic cells (**e**), and $CD11c^-Gr1^-CD11b^+F4/80^+$ TAMs (**f**) were determined by flow cytometry. **g**, Arg1 expression was determined by IHC. Representative images and quantitative data are shown ($n = 5$ per group). **h**, CD206 expression in TAMs was assessed by flow cytometry. **i**, Correlation between high and low tertiles of *RIP1–RIP3* co-expression and *CD11b* expression was tested in human PDA tissues using the UCSC RNA-seq database. Each point represents data from one patient. Graphs show mean ± s.e.m. $^*P < 0.05$, $^{**}P < 0.01$, $^{****}P < 0.0001$ (unpaired *t*-test). Flow cytometry experiments were carried out twice.



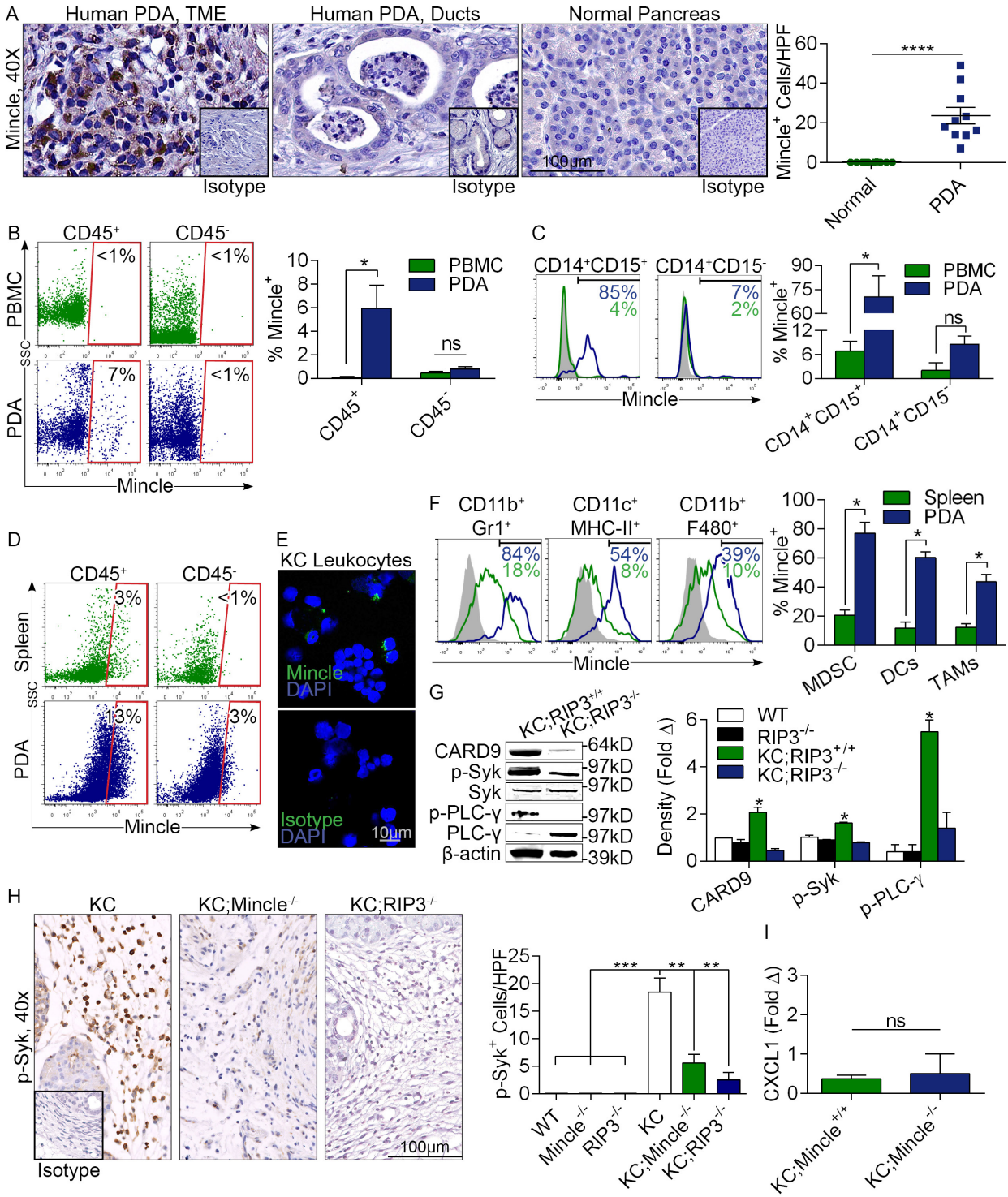
Extended Data Figure 3 | RIP3 deletion in PDA is associated with CD4$^+$ and CD8$^+$ T cell activation but RIP3 deletion does not alter growth of B16 melanomas or subcutaneously implanted pancreatic tumours. Wild-type and *Rip3*$^{-/-}$mice ($n = 7$ per group) were orthotopically implanted with KPC-derived tumour cells. **a–c**, Mice were killed three weeks later and intra-tumoral CD4$^+$ and CD8$^+$ T cell expression of IL-10 (**a**), PD-1 (**b**), and CD44 (**c**) was measured by flow cytometry. **d**, Co-expression of CD25 and FoxP3 on CD4$^+$ T cells was also analysed. $^*P < 0.05$, $^{**}P < 0.01$ (unpaired *t*-test). Data were reproduced in two separate experiments. **e**, Wild-type and *Rip3*$^{-/-}$mice ($n = 3$ per group) were implanted subcutaneously with B16 melanoma cells and tumour size was measured at 4–7-day intervals. P = not significant at all time points. **f**, Wild-type and *Rip3*$^{-/-}$mice ($n = 3$ per group) were implanted subcutaneously with KPC-derived tumour cells and tumour size was measured at 4–7-day intervals. P values were not significant at all time points (unpaired *t*-test). Graphs show mean ± s.e.m.



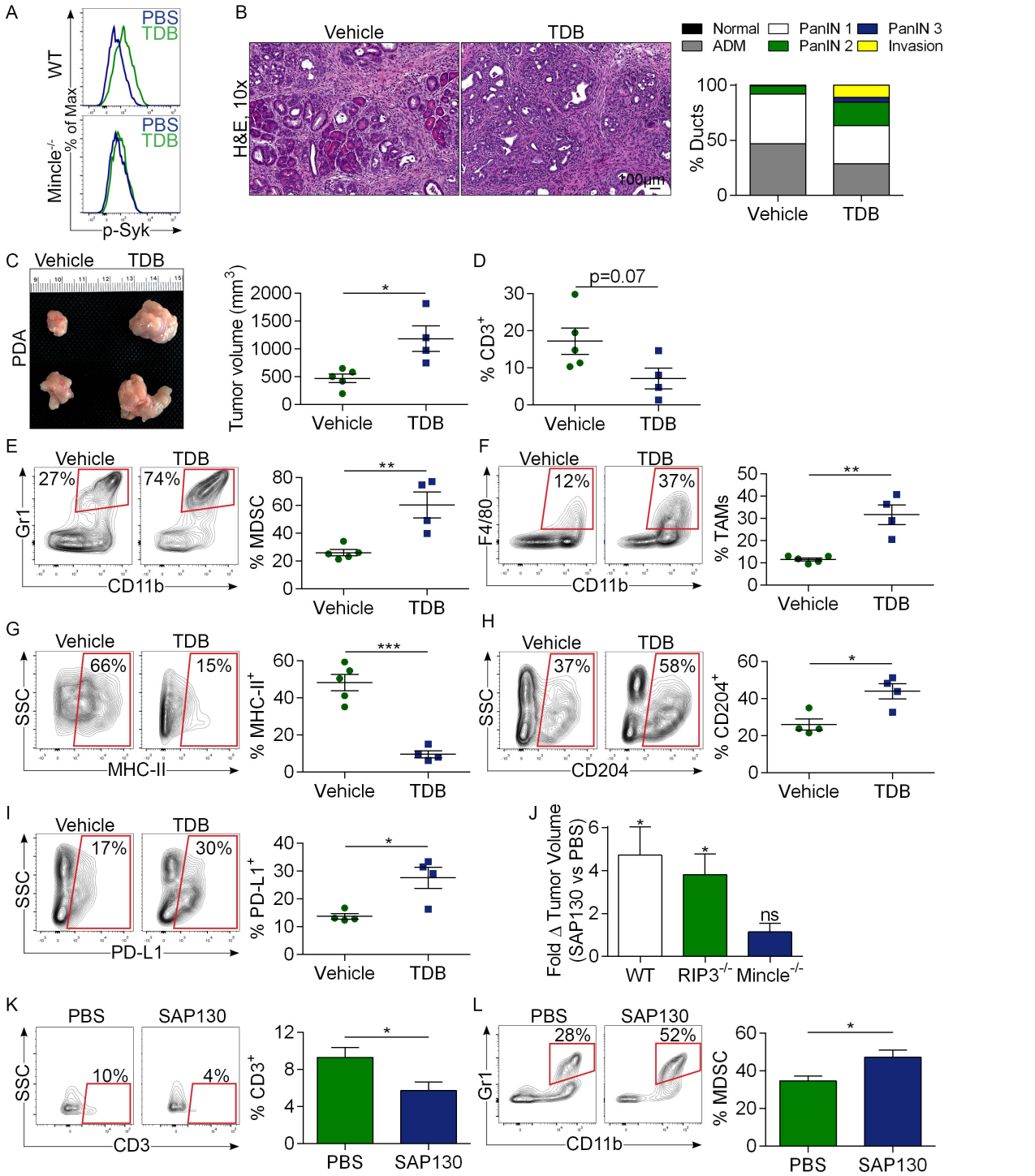
Extended Data Figure 4 | CXCL1 blockade protects against pancreatic oncogenesis. a, **b**, Pancreases from 6-month-old KC mice were analysed for co-expression of CD45 and CXCR2 (**a**) and CD68 and CXCR2 (**b**) by confocal microscopy. **c**, Wild-type mice were challenged with an orthotopic injection of KPC-derived tumour cells. Cohorts were treated thrice weekly with anti-CXCL1 monoclonal antibodies or isotype control. Pancreatic tumours were removed three weeks after implantation. Representative photographs and quantitative analyses of tumour volume and weight are shown ($n = 7$ per group). **d**, Wild-type ($n = 5$) and *Rip3*$^{-/-}$mice were challenged with an orthotopic injection of KPC-derived tumour cells. *Rip3*$^{-/-}$mice were serially treated with anti-CXCL1 monoclonal antibodies ($n = 3$) or isotype control ($n = 5$). Pancreatic tumours were harvested three weeks after implantation and tumour weight was recorded. **e**–**i**, Wild-type mice were challenged with an orthotopic injection of KPC-derived tumour cells and cohorts were serially treated with anti-CXCL1 monoclonal antibodies or isotype control. The fraction of peri-tumoral Gr1$^+$CD11b$^+$ MDSC (**e**) and Gr1$^-$CD11b$^+$F4/80$^+$ TAMs (**f**), the expression of CD44 (**g**) and TNFα (**h**) on CD3$^+$ T cells, and the fraction of peri-tumoral CD3$^+$ T cells (**i**) were determined by flow cytometry. Graphs show mean ± s.e.m. ns, not significant; $*P < 0.05$, $**P < 0.01$, $****P < 0.0001$ (unpaired *t*-test). Flow cytometry data were reproduced three times.



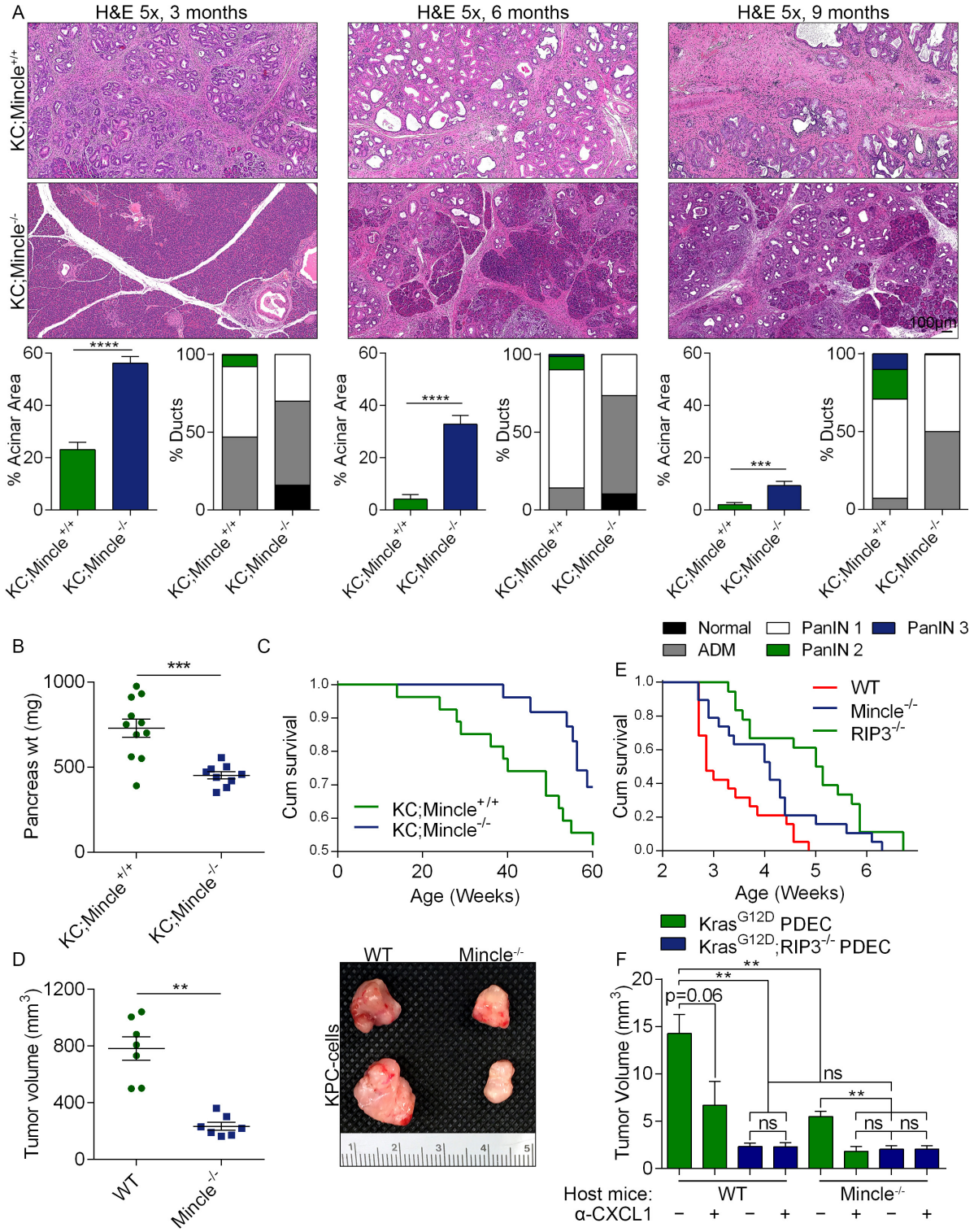
Extended Data Figure 5 | High SAP130 levels in PDA. a, Paraffin-embedded sections of human PDA and surrounding normal pancreas tissue were tested for expression of SAP130 by IHC compared with isotype control. Representative images and summary data from ten patients with PDA are shown. **b**, *SAP130* expression was measured by qPCR in human AsPC-1 cells after treatment with PBS or gemcitabine ($n = 3$ per group). **c**, SAP130 expression was assayed by IHC in paraffin-embedded pancreases from 6-month-old $p48^{Cre};Kras^{G12D};Rip3^{+/+}$, $p48^{Cre};Kras^{G12D};Rip3^{-/-}$, and wild-type mice ($n = 4$ per group) compared with respective isotype controls. Representative images and quantitative data are shown. **d**, *Sap130* expression was tested by qPCR in KPC-derived tumour cells treated with PBS or gemcitabine with or without Nec-1s in triplicate. **e**, **f**, SAP130 expression was tested by confocal microscopy in $CK19^{+}$ epithelial cells (**e**) and $CD45^{+}$ inflammatory cells (**f**) in mouse PDA. **g**, Co-expression of SAP130, RIP1, and RIP3 was tested by confocal microscopy in human PDA. **h**, Correlation between high and low tertiles of combined *RIP1/RIP3* and *SAP130* expression was tested using the UCSC RNA-seq database. Graphs show mean ± s.e.m. $^{*}P < 0.05$, $^{****}P < 0.0001$ (unpaired *t*-test). **i**, Patients with PDA with high or low tertile levels of *SAP130* expression were compared in a Kaplan–Meier survival analysis using the UCSC RNA-seq database. **j**, Pancreas lysate from 6-month-old wild-type or KC mice was immunoprecipitated using an anti-SAP130 or control antibody and then tested for expression of SAP130 and Mincle. Input controls were similarly probed. Results were reproduced in two separate experiments. For gel source data, see Supplementary Fig. 1.



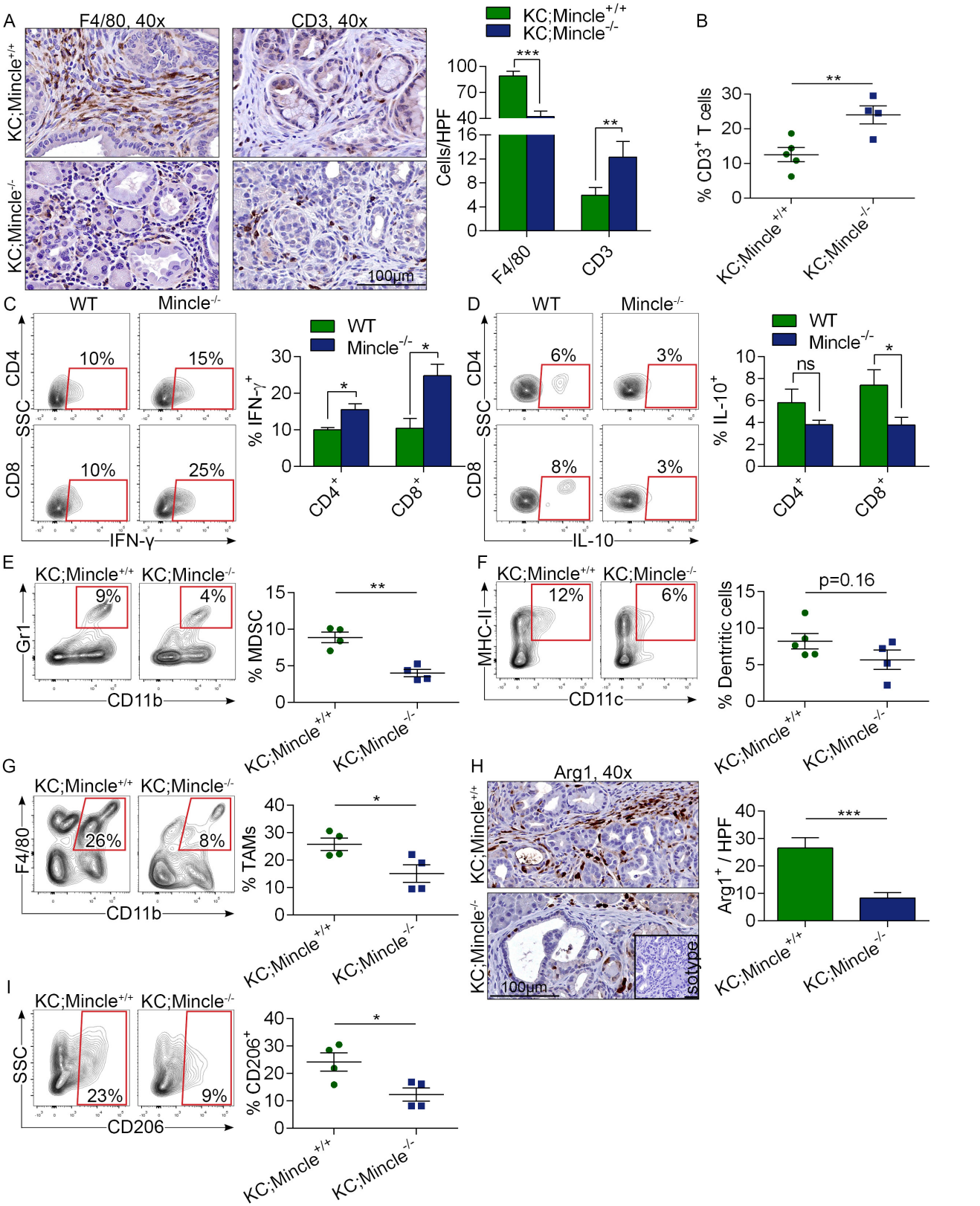
Extended Data Figure 6 | High Mincle signalling in PDA. a, Mincle expression was tested in paraffin-embedded sections of PDA and surrounding normal pancreas from ten patients with PDA. Representative stromal and ductal areas of PDA tumours and quantitative data are shown. **b**, $CD45^+$ and pancreas-infiltrating leukocytes and $CD45^-$ tumour or parenchymal cells from human PDA and PBMC were tested for Mincle expression and compared. **c**, PDA-infiltrating and PBMC-derived $CD14^+CD15^+$ and $CD14^+CD15^-$ cells from patients with PDA were gated and tested for Mincle expression compared with isotype control. Representative histograms and quantitative data are shown. **d**, $CD45^+$ and $CD45^-$ cells from PDA and spleens from 6-month-old KC mice were tested for expression of Mincle. Representative histograms are shown. **e**, Pancreas-infiltrating leukocyte suspensions from 6-month-old KC mice were tested for Mincle expression by immunofluorescence microscopy and compared with isotype control. **f**, Granulocytes, dendritic cells, and macrophages from PDA and spleens from 3-month-old KC mice were gated by flow cytometry and tested for expression of Mincle, and compared with isotype control. Representative histograms and quantitative data are shown ($n = 3$). **g**, Whole pancreas lysates from wild-type, *Rip3*$^{-/-}$, *p48*^{Cre};*Kras*^{G12D};*Rip3*$^{+/+}$, and *p48*^{Cre};*Kras*^{G12D};*Rip3*$^{-/-}$ mice were probed for CARD9, p-Syk, Syk, p-PLC-γ, and PLC-γ by western blotting. Density analysis was performed in triplicate. **h**, Pancreases from wild-type, *Mincle*$^{-/-}$, *Rip3*$^{-/-}$, *p48*^{Cre};*Kras*^{G12D}, *p48*^{Cre};*Kras*^{G12D};*Mincle*$^{-/-}$, and *p48*^{Cre};*Kras*^{G12D};*Rip3*$^{-/-}$ mice ($n = 3$ per group) were stained using a monoclonal antibody directed against p-Syk. Representative images and quantitative data are shown. Graphs show mean ± s.e.m. ns, not significant; $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$ (unpaired *t*-test). **i**, Pancreases from 3-month-old KC and *p48*^{Cre};*Kras*^{G12D};*Mincle*$^{-/-}$ mice were tested for CXCL1 expression by PCR in biological duplicates. Data were reproduced twice. For gel source data, see Supplementary Fig. 1.



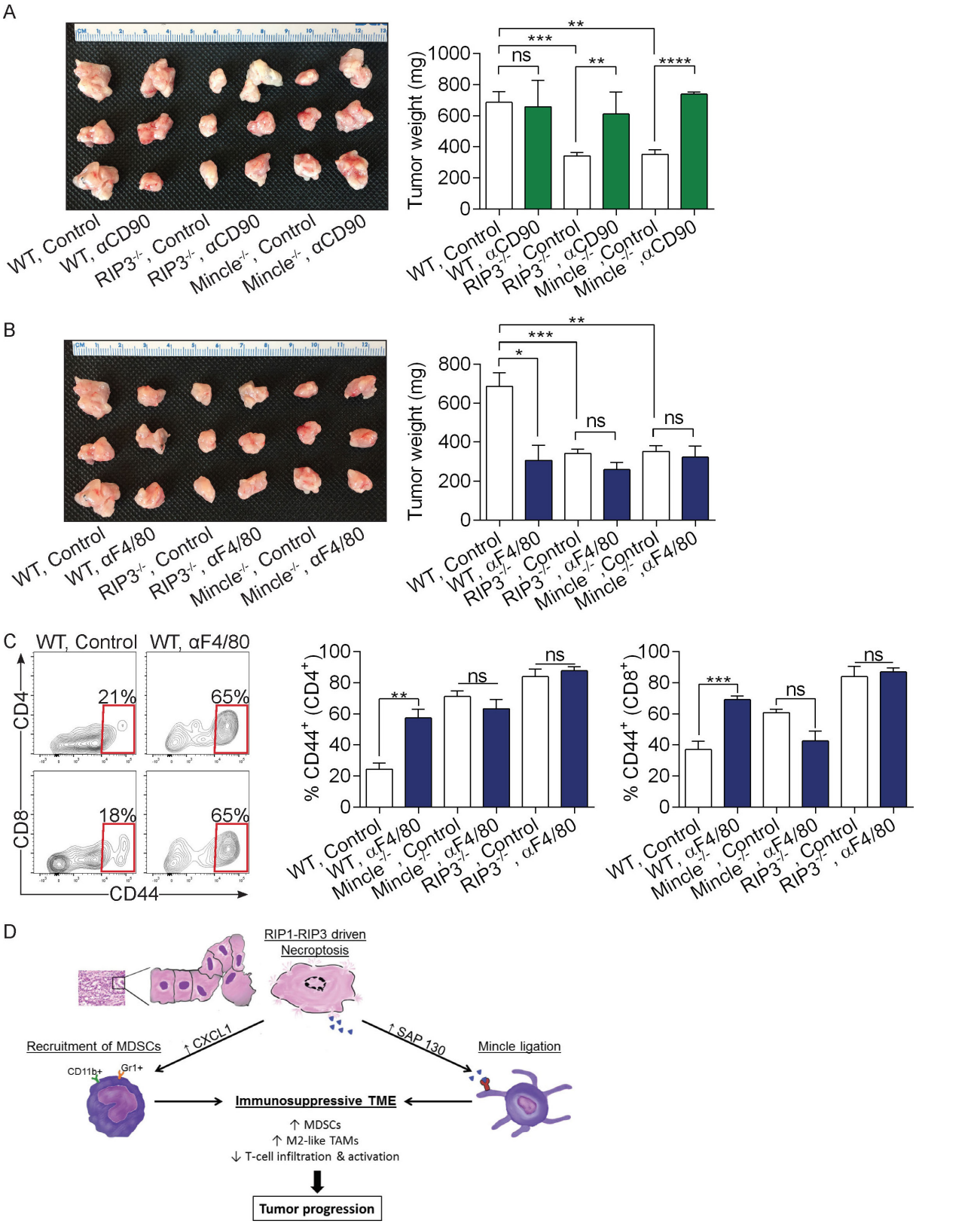
Extended Data Figure 7 | Mincle ligation accelerates pancreatic oncogenesis. a, Wild-type or *Mincle*$^{-/-}$ mice ($n = 4$ per group) were administered a single dose of vehicle or the Mincle ligand TDB and p-Syk expression was tested in Gr1^{-}F4/80^{+}CD11b^{+} splenic macrophages 4 h later by flow cytometry. Representative data are shown. **b**, Six-week-old KC mice were treated with TDB or vehicle for 8 weeks before being killed ($n = 5$ per group). Representative H&E-stained sections are shown and the fraction of ducts exhibiting normal morphology, ADM, graded PanIN lesions, or foci of invasive cancer are shown. **c–i**, *Rip3*$^{-/-}$ mice were orthotopically implanted with KPC-derived tumour cells and treated thrice weekly with TDB ($n = 4$) or vehicle ($n = 5$) before being killed 3 weeks after implantation. **c**, Representative images of tumours and pancreatic weights and tumour volume are shown. **d–i**, The fraction of CD3^{+} T cells (**d**), Gr1^{+}CD11b^{+} MDSC (**e**), and Gr1^{-}CD11b^{+}F4/80^{+} TAMs (**f**) was determined by flow cytometry. Expression of MHC II (**g**), CD204 (**h**), and PD-L1 (**i**) in TAMs is shown for each cohort. Data were reproduced in two separate experiments. **j**, Wild-type ($n = 4$ per group), *Rip3*$^{-/-}$ ($n = 4$ per group), and *Mincle*$^{-/-}$ ($n = 3$ per group) mice were challenged orthotopically with KPC-derived tumour cells. On days 7 and 14 after implantation, mice underwent mini-laparotomies, tumour volume was measured *in situ*, and PBS or recombinant SAP130 was injected into the tumours. On day 20, mice were killed and the final tumour volume was recorded. The fold-increase in tumour volume between days 7 and 20 in SAP130- versus PBS-treated tumours is shown for wild-type, *Rip3*$^{-/-}$, and *Mincle*$^{-/-}$ mice. **k**, **l**, Wild-type mice were similarly challenged with orthotopic KPC-derived tumour and then given intra-tumoral injections of PBS ($n = 4$) or recombinant SAP130 ($n = 3$) on days 7 and 14. On day 20, tumours were removed and the fraction of CD3^{+} T cells (**k**) and Gr1^{+}CD11b^{+} MDSC (**l**) among CD45^{+} tumour-infiltrating leukocytes was determined. Graphs show mean ± s.e.m. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$ (unpaired *t*-test).



Extended Data Figure 8 | Mincle deletion protects against pancreatic oncogenesis. a, $p48^{Cre}$;$Kras^{G12D}$;$Mincle^{+/+}$ ($n=11$) and $p48^{Cre}$;$Kras^{G12D}$;$Mincle^{-/-}$ ($n=9$) mice were killed at 3, 6, or 9 months of age. Representative H&E-stained sections of pancreases are shown. The percentage of pancreas occupied by intact acinar structures, and the fractions of ducts exhibiting normal morphology, ADM, or graded PanIN I–III lesions were calculated. **b**, Weights of pancreases from 3-month-old $p48^{Cre}$;$Kras^{G12D}$;$Mincle^{+/+}$ ($n=11$) and $p48^{Cre}$;$Kras^{G12D}$;$Mincle^{-/-}$ ($n=9$) mice. **c**, Kaplan–Meier survival analysis was performed for $p48^{Cre}$;$Kras^{G12D}$;$Mincle^{+/+}$ ($n=29$) and $p48^{Cre}$;$Kras^{G12D}$;$Mincle^{-/-}$ ($n=28$) mice ($P=0.06$). The controls were the same as for the experiments shown in Fig. 3. **d**, KPC-derived tumour cells were orthotopically implanted in the pancreases of wild-type or $Mincle^{-/-}$ mice. Animals were killed 3 weeks after implantation ($n=7$ per group). Tumour volume was recorded. Representative images of pancreatic tumours are shown. **e**, KPC-derived tumour cells were orthotopically implanted in the pancreases of wild-type ($n=19$), $Mincle^{-/-}$ ($n=19$), and $Rip3^{-/-}$ ($n=18$) mice. Kaplan–Meier survival analysis was performed (wild-type vs $Mincle^{-/-}$: $P=0.03$; wild-type vs $Rip3^{-/-}$: $P<0.0001$; $Mincle^{-/-}$ vs $Rip3^{-/-}$: $P=0.03$). **f**, Wild-type and $Mincle^{-/-}$ mice were orthotopically implanted with $Kras^{G12D}$;$Rip3^{+/+}$ PDEC or $Kras^{G12D}$;$Rip3^{-/-}$ PDEC. Mice were treated with a neutralizing anti-CXCL1 monoclonal antibody or isotype control (mean $n=4$ per group). Mice were killed 3 weeks after implantation and tumour volume was recorded. Graphs show mean ± s.e.m. ns, not significant; $^{**}P<0.01$, $^{***}P<0.001$, $^{****}P<0.0001$ (unpaired t-test).



Extended Data Figure 9 | Mincle deletion in PDA enhances the immunogenicity of the inflammatory TME. a, $p48^{Cre};Kras^{G12D};Mincle^{+/+}$ and $p48^{Cre};Kras^{G12D};Mincle^{-/-}$ mice were killed at 3 months of age. Paraffin-embedded sections of pancreas were stained using monoclonal antibodies directed against F4/80 and CD3 ($n = 5$ per group). Representative images and quantitative data are shown. **b**–**d**, The fraction of peri-tumoral $CD3^+$ T cells (**b**) and expression of IFN-γ (**c**) and IL-10 (**d**) on $CD4^+$ and $CD8^+$ T cells were determined by flow cytometry. **e**–**g**, The fraction of tumour-infiltrating $Gr1^+CD11b^+$ MDSC (**e**), $F4/80^-CD11c^+MHC\,II^+$ dendritic cells (**f**), and $Gr1^-CD11b^+F4/80^+$ TAMs (**g**) was also determined by flow cytometry. **h**, Arg1 expression was determined by IHC. Representative images and quantitative data are shown. **i**, CD206 expression in TAMs was determined by flow cytometry. Graphs show mean ± s.e.m. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$ (unpaired *t*-test). Experiments were performed twice with similar results.



Extended Data Figure 10 | RIP3 and Mincle signalling are necessary for macrophage-induced suppression of T cell immunity in PDA. **a**–**c**, Wild-type, *Rip3*$^{-/-}$, and *Mincle*$^{-/-}$ mice were challenged with orthotopic implantation of PDA cells. Before tumour implantation, mice were treated daily for 3 days with a neutralizing anti-CD90 monoclonal antibody (**a**), a neutralizing anti-F4/80 monoclonal antibody (**b**), or isotype control. Antibodies were administered twice weekly for the duration of the experiment. Mice were killed 21 days after implantation and the pancreatic tumours were weighed. Controls were shared for both experiments and are shown twice ($n = 4$ for *Mincle*$^{-/-}$ anti-CD90 and anti-F4/80-treated groups; $n = 3$ for other groups). (**c**) CD4$^+$ and CD8$^+$ T cell activation was determined by expression of CD44 in wild-type, *Rip3*$^{-/-}$, and *Mincle*$^{-/-}$ mice treated with anti-F4/80 monoclonal antibody or isotype control. Graphs show mean ± s.e.m. ns, not significant; $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.001$ (unpaired *t*-test). *In vivo* cellular depletion experiments were performed on two separate occasions with similar results. **d**, Schematic depicting immunosuppressive implications of RIP1/RIP3-driven CXCL1 expression and Mincle activation.

LETTER

doi:10.1038/nature17392

sFRP2 in the aged microenvironment drives melanoma metastasis and therapy resistance

Amanpreet Kaur[1,2], Marie R. Webster[1], Katie Marchbank[1], Reeti Behera[1], Abibatou Ndoye[1], Curtis H. Kugel III[1], Vanessa M. Dang[1], Jessica Appleton[1], Michael P. O'Connell[1], Phil Cheng[3], Alexander A. Valiga[1], Rachel Morissette[4], Nazli B. McDonnell[4], Luigi Ferrucci[4], Andrew V. Kossenkov[1], Katrina Meeth[5], Hsin-Yao Tang[1], Xiangfan Yin[1], William H. Wood III[4], Elin Lehrmann[4], Kevin G. Becker[4], Keith T. Flaherty[6], Dennie T. Frederick[6], Jennifer A. Wargo[7], Zachary A. Cooper[7], Michael T. Tetzlaff[7], Courtney Hudgens[7], Katherine M. Aird[1], Rugang Zhang[1], Xiaowei Xu[8], Qin Liu[1], Edmund Bartlett[8], Giorgos Karakousis[8], Zeynep Eroglu[9], Roger S. Lo[10], Matthew Chan[11], Alexander M. Menzies[12], Georgina V. Long[12], Douglas B. Johnson[13], Jeffrey Sosman[13], Bastian Schilling[14,15], Dirk Schadendorf[14,15], David W. Speicher[1], Marcus Bosenberg[5], Antoni Ribas[10] & Ashani T. Weeraratna[1]

Cancer is a disease of ageing. Clinically, aged cancer patients tend to have a poorer prognosis than young. This may be due to accumulated cellular damage, decreases in adaptive immunity, and chronic inflammation. However, the effects of the aged microenvironment on tumour progression have been largely unexplored. Since dermal fibroblasts can have profound impacts on melanoma progression[1–4], we examined whether age-related changes in dermal fibroblasts could drive melanoma metastasis and response to targeted therapy. Here we find that aged fibroblasts secrete a Wnt antagonist, sFRP2, which activates a multi-step signalling cascade in melanoma cells that results in a decrease in β-catenin and microphthalmia-associated transcription factor (MITF), and ultimately the loss of a key redox effector, APE1. Loss of APE1 attenuates the response of melanoma cells to DNA damage induced by reactive oxygen species, rendering the cells more resistant to targeted therapy (vemurafenib). Age-related increases in sFRP2 also augment both angiogenesis and metastasis of melanoma cells. These data provide an integrated view of how fibroblasts in the aged microenvironment contribute to tumour progression, offering new possibilities for the design of therapy for the elderly.

To determine whether the aged microenvironment promotes melanoma progression, Yumm1.7 cells, derived from the $Braf^{V600E}$/$Cdkn2a^{-/-}$/$Pten^{-/-}$ mouse model of melanoma[5], were injected into 8-week-old mice (young) or 52-week-old mice (aged). Tumours grew slowly in aged mice (Fig. 1a), but were more aggressive, with increased angiogenesis (Fig. 1b) and lung metastases (Fig. 1c and Extended Data Fig.1a). This is consistent with observations that melanoma cells switch between proliferative and invasive states, termed 'phenotype switching', which depends on changes in Wnt signalling and MITF[6–8]. We have shown that phenotype switching to a metastatic, therapy-resistant state can be induced by stresses such as hypoxia[9].

Cross-talk between dermal fibroblasts and transformed melanocytes is critical for melanoma invasion[1–4] and therapy resistance[10]. Dermal fibroblasts senesce during ageing[11], and senescent fibroblasts can promote tumour invasion[12]. To determine whether fibroblasts from normally aged skin could promote tumour progression, we used dermal fibroblasts from young (<35 years) and aged (>55 years) healthy donors from the Baltimore Longitudinal Study of Aging[13] (Extended Data Fig. 1). Proliferation was initially unaffected (Extended Data Fig. 1b), but aged fibroblasts senesced more rapidly than young fibroblasts (Extended Data Fig. 1c). Organotypic skin reconstructs were built using two melanoma cell lines (WM35 and WM793) and three young or three aged fibroblast lines, which persisted equally in the skin reconstructs as demonstrated by smooth muscle actin staining (Extended Data Fig. 1d). Melanoma cell invasion was increased, but proliferation was decreased in reconstructs built with aged fibroblasts (Fig. 1d, e and Extended Data Fig. 1e). To assess the effects of secreted factors, melanoma cells were exposed to conditioned media from aged or young fibroblasts. Aged fibroblast media decreased melanoma cell proliferation (Fig. 1f and Extended Data Fig. 1f), and increased invasion as measured by spheroid and Boyden chamber invasion assays (Fig. 1g, h and Extended Data Fig. 1g). These data indicate that secreted factors from fibroblasts play critical roles in phenotype switching in melanoma cells.

A key player in phenotype switching is β-catenin, which drives proliferation and inhibits invasion of melanoma cells[14,15]. An inhibitor of β-catenin, sFRP2, was significantly increased in the secretome of aged fibroblasts (Fig 2a). In both spheroid and Boyden chamber assays, media from young fibroblasts treated with recombinant sFRP2 (rsFRP2) increased invasion of melanoma cells (Fig. 2b and Extended Data Fig. 2a) and media from aged fibroblasts treated with α-sFRP2 inhibited melanoma cell invasion (Fig. 2c and Extended Data Fig. 2c). *In vivo*, sFRP2 was increased in sera from aged mice, and these levels could be nearly attained in young mice by intravenous injection of rsFRP2 (Fig. 2d). Increasing serum sFRP2 in young mice increased metastasis of Yumm1.7 to the lung (Fig. 2e) and increased angiogenesis in subcutaneous tumours (Fig. 2f and Extended Data Fig. 2c, d). Tumours in α-sFRP2-treated mice showed impaired angiogenesis (Fig. 2f and Extended Data Fig. 2e, f). However, α-sFRP2-treated aged mice succumbed to a lethal inflammation (Extended Data Fig. 3), an outcome previously observed in aged immune-competent mice treated with α-PD1 (ref. 16).

We next explored the significance of sFRP2 effects on β-catenin expression and activity. Treatment of melanoma cells with rsFRP2 inhibited β-catenin expression (Fig. 2g), and β-catenin was decreased in aged human skin (Fig. 2h and Extended Data Fig. 4a) and in melanoma cells treated with aged fibroblast media (Extended Data Fig. 4b, c). Yumm1.7 tumours in aged mice exhibited decreased β-catenin and increased sFRP2

[1]The Wistar Institute, Philadelphia, Pennsylvania 19104, USA. [2]University of the Sciences, Philadelphia, Pennsylvania 19104, USA. [3]Department of Dermatology, University of Zurich, Zurich CH-8006, Switzerland. [4]The National Institute on Aging, National Institutes of Health, Baltimore, Maryland 21224, USA. [5]Department of Dermatology and Pathology, Yale University, New Haven, Connecticut 06511, USA. [6]Massachusetts General Hospital Cancer Center, Developmental Therapeutics, Boston 02114, Massachusetts, USA. [7]Department of Surgical Oncology, MD Anderson Cancer Center, Houston, Texas 77030, USA. [8]Departments of Surgery and Pathology, Abramson Cancer Center, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [9]Department of Medical Oncology, City of Hope Medical Center, Duarte, California 91010, USA. [10]Department of Medicine, Division of Hematology-Oncology, University of California Los Angeles, Los Angeles, California 90095, USA. [11]Crown Princess Mary Cancer Centre, Westmead Hospital, Westmead 2145, Australia. [12]Melanoma Institute Australia and The University of Sydney, Sydney 2000, Australia. [13]Department of Medicine, Vanderbilt University Medical Center, Nashville Tennessee 37232, USA. [14]Department of Dermatology, University Hospital, West German Cancer Center, University Duesburg-Essen, Essen, Germany. [15]German Cancer Consortium (DKTK), Heidelberg 45127, Germany.

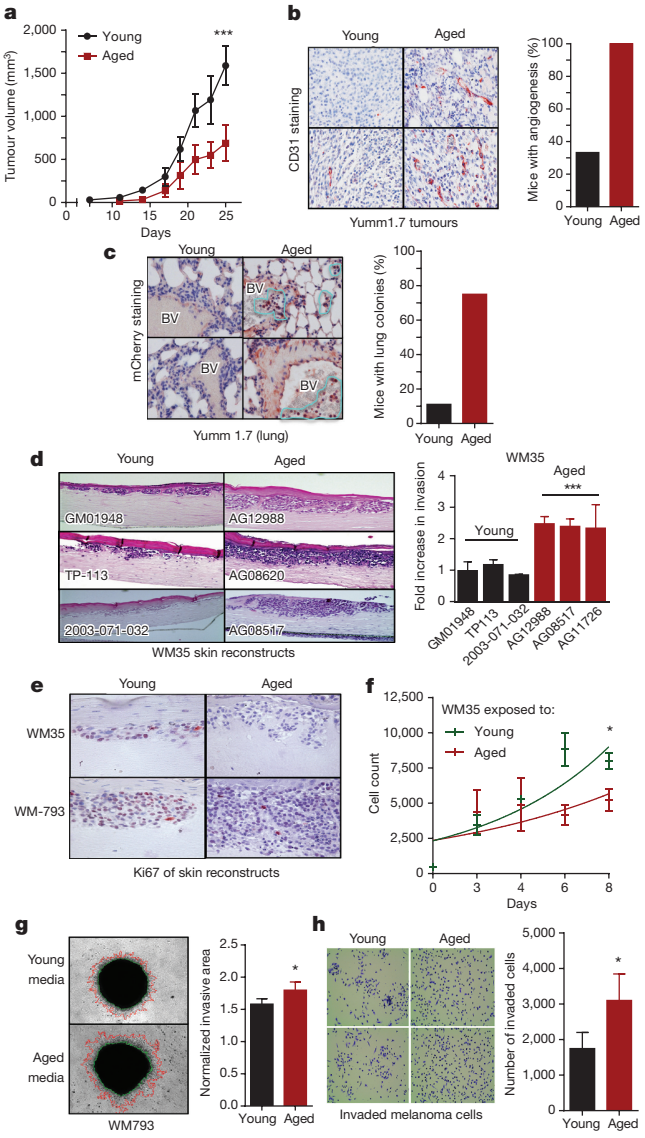


Figure 1 | The aged microenvironment promotes phenotype switching. **a**, Growth of 1×10^6 YUMM1.7_mCherry cells subcutaneously injected in young (8 weeks, $n = 10$) and aged (52 weeks, $n = 10$) C57/BL6 mice. (Analysis of variance (ANOVA) $P < 0.0001$; multiple comparison's test using Bonferroni correction, $P < 0.0001$ after day 19.) **b**, CD31 expression from **a** (original magnification ×200). Graph indicates percentage of mice with extensive angiogenesis (multiple $CD31^+$ vessels per field). **c**, mCherry staining in mouse lungs (positive cells circled in blue; original magnification ×400). **d**, WM35 melanoma cells in skin reconstructs built with young or aged fibroblasts ($n = 3$, original magnification ×150). Invasion was quantified using ImageJ (ANOVA $P = 0.0006$). **e**, Ki67 staining from **d** (original magnification ×400). **f**, Proliferation of WM35 melanoma cells in conditioned media from young or aged fibroblasts ($n = 4$ fibroblasts per group; ANOVA, $P = 0.004$; Holm–Šídák correction (day 6 ($P = 0.002$), day 8 ($P = 0.034$)). **g**, Invasion of melanoma spheroids exposed to aged or young conditioned media ($n = 2$ fibroblasts per group, original magnification ×40, two-tailed unpaired t-test, $P = 0.041$). **h**, Boyden chamber invasion of FS5 melanoma cells treated with young versus aged fibroblast media (original magnification ×150, two-tailed unpaired t-test, $P = 0.043$). Data represented as mean ± s.d. (**a**, **d**, **g**, **h**).

(Fig. 2i), effects reversed by treatment of aged mice with α-sFRP2 (Fig. 2j). Knockdown of β-catenin in young fibroblasts increased their sFRP2 secretion (Fig. 2k), which reduced β-catenin and increased invasion in melanoma cells (Fig. 2l and Extended Data Fig. 4d). These data suggest that sFRP2 present in the aged microenvironment may contribute to melanoma metastasis through suppression of β-catenin in melanoma cells.

The loss of β-catenin permits oxidative stress in haematopoietic stem cells[17]. We hypothesized that the same might occur in melanoma cells, via MITF, which is activated by β-catenin[18]. MITF modulates responses to reactive oxygen species (ROS) by increasing the redox effector APE1 (REF-1)[19]. We predicted that age-induced loss of β-catenin might dysregulate the MITF/APE1 redox pathway. Indeed, melanoma cells exposed to aged fibroblast media have decreased β-catenin, MITF, and APE1 (Fig. 3a). APE1 expression is lower in aged human skin (Extended Data Fig. 5a) and in Yumm1.7 tumours in aged mice (Fig. 3b). Loss of APE1 reduces cellular responses to ROS. ROS activity is increased in aged fibroblasts (Fig. 3c), partly because aged fibroblasts secrete lower levels of superoxide dismutase 3 (SOD3), and peroxiredoxin 6, which scavenge free radicals (Extended Data Fig. 5b, c). A marker of oxidative stress, 8-oxo-dG, was increased in aged but not young skin (Extended Data Fig. 5d) and in tumours in aged mice (Fig. 3d). Knockdown of APE1 in melanoma cells increases ROS activity, especially upon exposure to aged fibroblast media (Fig. 3e, f and Extended Data Fig. 5e). We hypothesized that this loss of ability to respond to ROS in an aged environment is partly due to sFRP2 downregulation of APE1. Treating young mice with rsFRP2 increases 8-oxo-dG in Yumm1.7 tumours (Fig. 3g), while depleting sFRP2 in aged mice decreases 8-oxo-dG (Fig. 3g). Together, these data indicate that sFRP2 increases oxidative stress by inhibiting β-catenin and, ultimately, APE1 in melanoma cells (Fig. 3h).

Our model indicates that loss of APE1 owing to age-induced sFRP2 should render melanoma cells sensitive to DNA damage (Extended Data Fig. 6a). Accordingly, microarray analysis reveals an increased DNA damage response signature in melanoma cells exposed to aged fibroblasts (Extended Data Fig. 6b, c). The DNA damage markers γH2AX and 53BP1 are increased in melanoma cells grown in aged mice (Fig. 4a), or exposed to aged fibroblast media (Fig. 4b) or in skin reconstructs built with aged fibroblasts (Extended Data Fig. 7a, b). Melanoma cells exposed to aged media also had increased levels of DNA damage (Fig. 4c and Extended Data Fig. 7c). Pre-treating aged fibroblasts with the anti-oxidant *N*-acetyl-L-cysteine (NAC) decreased ROS (Extended Data Fig. 7d) and, subsequently, γH2AX in melanoma cells (Fig. 4d). Conversely, knockdown of the anti-oxidant SOD3 in young fibroblasts increased γH2AX in melanoma cells (Fig. 4e and Extended Data Fig. 7e), suggesting that age-induced DNA damage could be reversed by inhibiting ROS activity.

We next queried the contribution of the sFRP2 → APE1 signalling axis to ROS-induced DNA damage. Loss of APE1 increases γH2AX (Fig. 4f). In melanoma cells, γH2AX also increases with treatment of rsFRP2, and decreases in cells where sFRP2 is deleted (Fig. 4g and Extended Data Fig. 7f, g). *In vivo*, 53BP1 increases in Yumm1.7 tumours in young mice treated with rsFRP2, and decreases in aged mice treated with α-sFRP2 (Fig. 4h). Aged melanoma patients (over 55 years) had significantly higher serum levels of sFRP2 ($P = 0.0084$) than young patients (under 40 years) (Fig. 4i). The key players in the pathway outlined in Fig. 3h were altered in aged melanoma patients: increased sFRP2 (in both fibroblasts and melanoma cells), decreased β-catenin, MITF, and APE1, and increased 8-oxo-dG and 53BP1 (Fig. 4j and Extended Data Fig. 8). These data confirmed our observations both from human cell line and from mouse data.

Increases in ROS[20,21] and decreases in β-catenin[9,22] and MITF[23] have been associated with increased resistance to BRAF inhibitors. Accordingly, spheroids treated with young medium were more sensitive to PLX4720 (vemurafenib) than those exposed to aged medium (Extended Data Fig. 9a). *In vivo*, Yumm1.7 tumours in young mice responded to PLX4720 more robustly than tumours in aged mice (Fig. 5a, b). Melanoma cells first exposed to media from young fibroblasts treated with H_2O_2 (to increase ROS) developed resistance to PLX4720 (Fig. 5c). However, when we first exposed melanoma cells to media from aged fibroblasts pre-treated with NAC, the melanoma cells died, whether treated with NAC alone or with NAC + PLX4720 (Fig. 5d, e). This indicated that vemurafenib-resistant cells in an aged microenvironment might be uniquely sensitive to anti-oxidants.

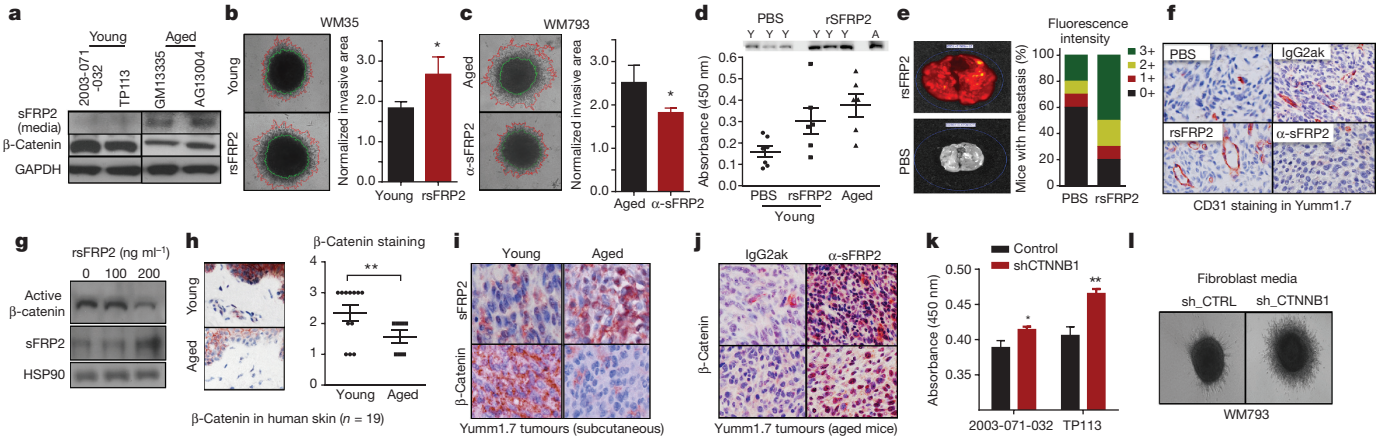


Figure 2 | sFRP2 promotes metastasis during ageing. a, Western analysis of sFRP2 (media) and β-catenin (cell lysates) in fibroblasts. **b**, Invasion of melanoma spheroids (original magnification ×40), treated with rsFRP2 (200 ng ml^{-1}, 48 h, two-tailed unpaired *t*-test, $P = 0.046$). **c**, Melanoma spheroids treated with media from aged fibroblasts pre-treated with either IgG2aκ, or α-sFRP2 monoclonal antibody (15 μg ml^{-1}, 48 h, original magnification ×40, two-tailed unpaired *t*-test, $P = 0.023$). **d**, sFRP2 in mouse serum ($n = 10$ per group): young, aged, and young + 200 ng ml^{-1} rsFRP2. ANOVA ($P = 0.015$), two-tailed unpaired *t*-test (PBS versus rsFRP2 ($P = 0.045$), young versus aged mice ($P = 0.007$)). **e**, One million mCherry-labelled Yumm1.7 cells were injected intravenously into PBS- or rsFRP2- (200 ng ml^{-1}) treated young mice (6–8 weeks, $n = 10$ per group). Graph indicates percentage of positive lungs recorded as IVIS fluorescence intensity where highest intensity is 3+ (green; that is, multiple metastatic foci) and 0 is lowest intensity (black; that is, no metastases). **f**, Yumm1.7 tumours in young mice (6–8 weeks, $n = 10$ per group) treated with either PBS or rsFRP2 (200 ng ml^{-1}) or in aged mice (52 weeks, $n = 5$ per group) treated with either control IgG2aκ, or α-sFRP2 monoclonal antibody (1 mg kg^{-1}) were assessed for CD31 (original magnification ×400). **g**, Western analysis of non-phosphorylated (active) β-catenin in melanoma cells treated with rsFRP2 for 48 h. **h**, β-Catenin expression in young and aged human skin (<35 years, $n = 12$; >55 years, $n = 7$). Slides were scored for positivity (3, highest; 0, lowest) (original magnification ×400, unpaired *t*-test using rank-sum ($P = 0.019$)). **i**, sFRP2 and β-catenin in Yumm1.7 tumours in young and aged mice (original magnification ×600). **j**, Aged tumour-bearing mice (>52 weeks, $n = 5$ per group) treated with α-sFRP2 monoclonal antibody (1 mg kg^{-1}). Tumours were stained for β-catenin (original magnification ×400). **k**, sFRP2 ELISA in shCTNNB1 versus shCTRL fibroblast media. Two-tailed unpaired *t*-test (2003-071-032 ($P = 0.015$), TP113 ($P = 0.008$)). **l**, Invasion of melanoma spheroids treated with TP113 shCTNNB1 conditioned media for 48 h (original magnification ×40). Data represented as mean ± s.d. (**b**–**d**, **h**, **k**).

This finding is especially exciting in light of recent data that indicate that anti-oxidants are effective in treating KRAS and BRAF mutant colon cancers[24].

Phenotype switching resulting in the loss of β-catenin has also been shown to contribute to resistance to vemurafenib[9,22]. β-Catenin expression is elevated in sensitive versus resistant cell lines (Extended Data Fig. 9b). Further, patients with clinical response to vemurafenib (over 30% tumour reduction as measured by RECIST) had higher levels of β-catenin than those with muted response (Fig. 5f, g). Yumm1.7 cells are sensitive to PLX4720, and knocking down *Ctnnb1* in these cells significantly increased their resistance to PLX4720 (Extended Data Fig. 9c). To determine whether sFRP2 alters the response of melanoma cells to vemurafenib, rsFRP2 was administered intravenously to tumour-bearing young mice, before and during PLX4720 exposure. β-Catenin was decreased in the skin of sFRP2-treated mice (Extended Data Fig. 9d), and treated mice rapidly developed resistance to PLX4720 (Fig. 5h). Depletion of sFRP2 in aged mice began to re-sensitize tumours to PLX4720 (Fig. 5i and Extended Data Fig. 9e), but a firm conclusion cannot be drawn from these data because of the premature death of the mice as described earlier.

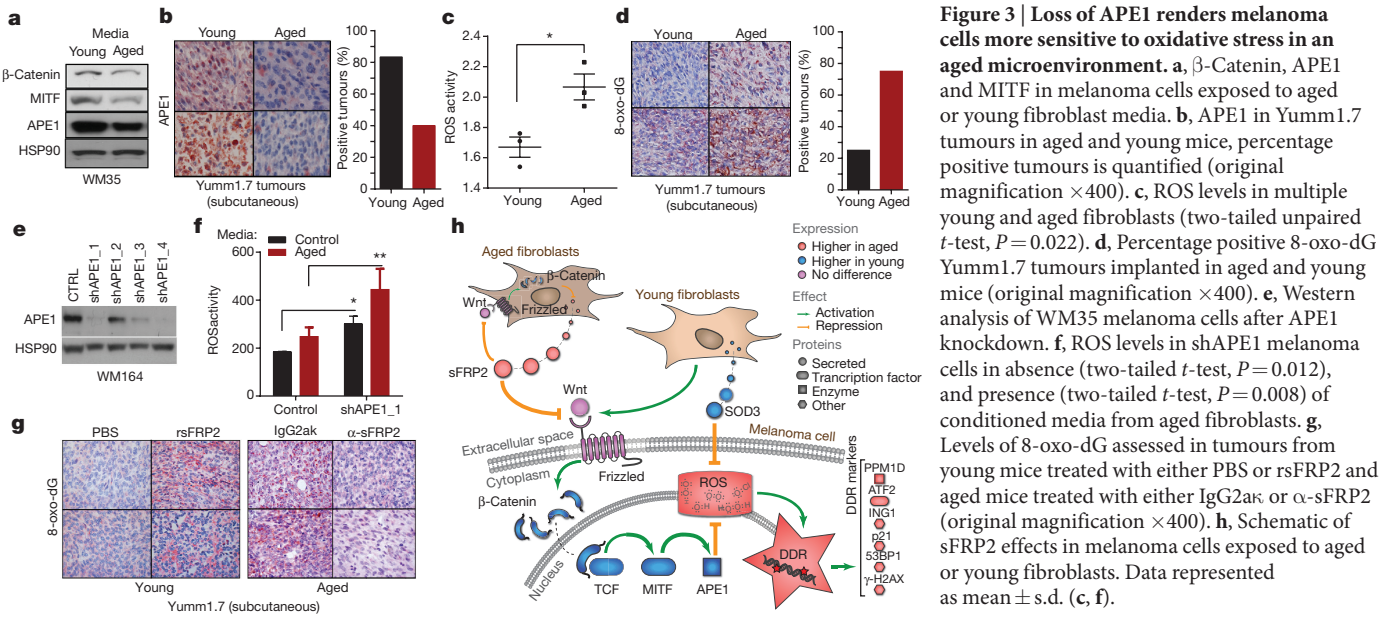


Figure 3 | Loss of APE1 renders melanoma cells more sensitive to oxidative stress in an aged microenvironment. a, β-Catenin, APE1 and MITF in melanoma cells exposed to aged or young fibroblast media. **b**, APE1 in Yumm1.7 tumours in aged and young mice, percentage positive tumours is quantified (original magnification ×400). **c**, ROS levels in multiple young and aged fibroblasts (two-tailed unpaired *t*-test, $P = 0.022$). **d**, Percentage positive 8-oxo-dG Yumm1.7 tumours implanted in aged and young mice (original magnification ×400). **e**, Western analysis of WM35 melanoma cells after APE1 knockdown. **f**, ROS levels in shAPE1 melanoma cells in absence (two-tailed *t*-test, $P = 0.012$), and presence (two-tailed *t*-test, $P = 0.008$) of conditioned media from aged fibroblasts. **g**, Levels of 8-oxo-dG assessed in tumours from young mice treated with either PBS or rsFRP2 and aged mice treated with either IgG2aκ or α-sFRP2 (original magnification ×400). **h**, Schematic of sFRP2 effects in melanoma cells exposed to aged or young fibroblasts. Data represented as mean ± s.d. (**c**, **f**).

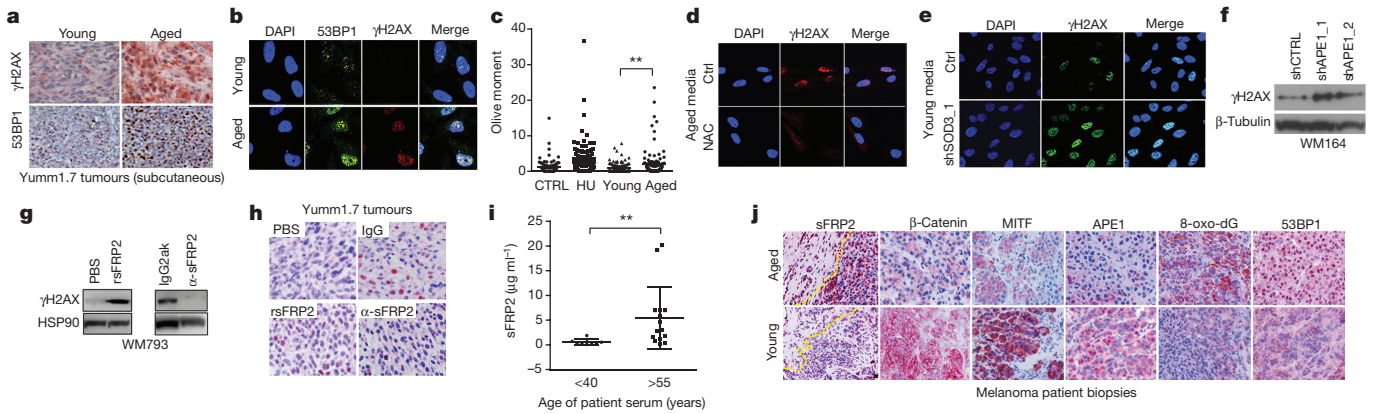


Figure 4 | Melanoma cells in an aged microenvironment exhibit increased DNA damage markers. a, Levels of γH2AX and 53BP1 in Yumm1.7 tumours in aged and young mice (original magnification ×400), and (**b**) in melanoma cells treated with conditioned media from young and aged fibroblasts. **c**, Comet assay analysis in melanoma cells treated with conditioned media from aged and young fibroblasts. Two-tailed unpaired *t*-test with Welch's correction (young versus aged, $P = 0.002$). Data represented as mean ± s.e.m. **d**, γH2AX in melanoma cells treated with conditioned media from aged fibroblasts pre-treated with 20 mM NAC (IF; 48 h), (**e**) conditioned medium from young fibroblasts with SOD3 knockdown (IF; 72 h), or (**f**) in melanoma cells with APE1 knockdown (Western blot; 48 h). **g**, γH2AX in melanoma cells exposed for 48 h to conditioned media from young fibroblasts pre-treated with rsFRP2 or aged fibroblasts treated with α-sFRP2 antibody (72 h). **h**, DNA damage marker 53BP1 in tumours from young mice treated with PBS or rsFRP2 and aged mice treated with IgG2aκ or α-sFRP2 (original magnification ×600). **i**, Serum sFRP2 ELISA in young and aged melanoma patients ($n = 8$ young, $n = 15$ aged; unpaired *t*-test with Welch's correction, $P = 0.008$). Data represented as mean ± s.d. **j**, Representative images from one young and aged patient for indicated proteins in the proposed pathway (original magnification ×400).

Since these data suggested that response to BRAF inhibitors might be attenuated in aged patients, we compared the age of patients ($n = 79$) at start of treatment with percentage tumour reduction (RECIST) after treatment with vemurafenib. Using age categories of under 35 and over 55, we saw a shift from a 41% (mean) response by RECIST in young patients to 25% (mean) response in patients over 55. Numbers of young patients were too low to achieve statistical significance. We therefore refocused our efforts on identifying an age cutoff where a statistically

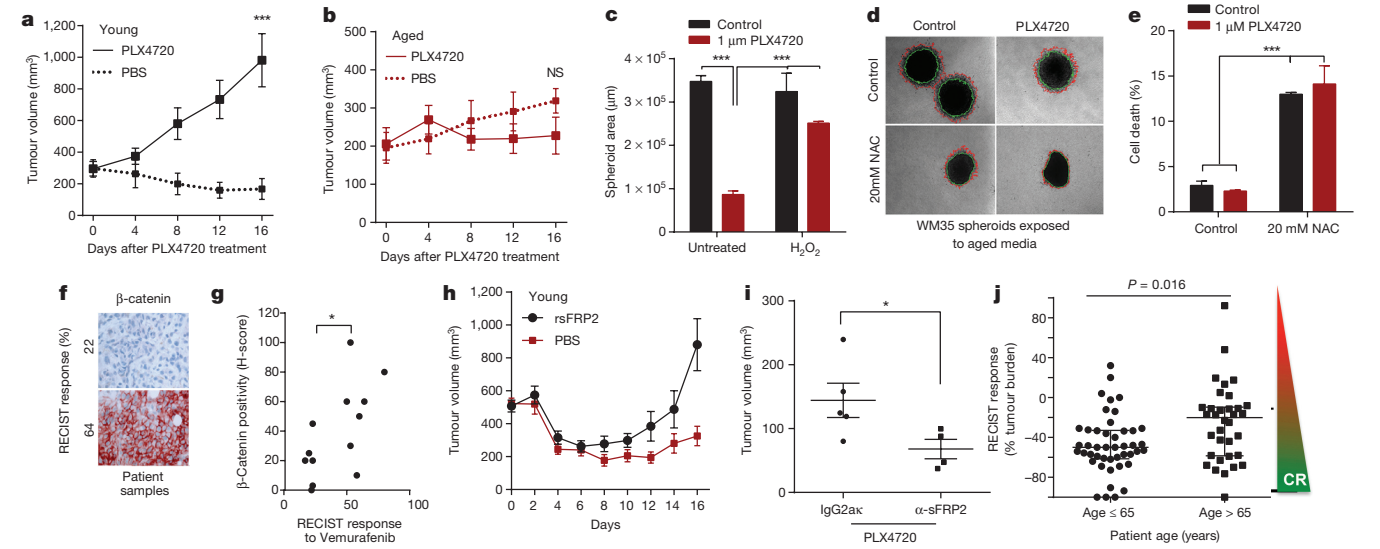


Figure 5 | The aged microenvironment induces therapy resistance. Yumm1.7 tumours in (**a**) young (8 weeks, $n = 10$ mice per treatment) or (**b**) aged (52 weeks, $n = 10$ mice per treatment) mice fed PLX4720 (417 mg kg^{-1}) chow. Tumours responded in (**a**) young ($P = 6 \times 10^{-5}$, two-way paired *t*-test), but not (**b**) aged mice ($P = 0.361$). Dotted lines indicate untreated controls. **c**, Melanoma spheroids were treated with 100 nM H_2O_2 (48 h), embedded in collagen and treated with young media containing 1 μM PLX4720 (ANOVA, $P < 0.0001$). Area was measured as a gauge of spheroid health. **d**, WM35 spheroids were embedded in collagen and treated with aged media containing 1 μM PLX4720 and/or 20 mM NAC. Invasion was measured after 48 h (representative images, original magnification ×40, one-way ANOVA, $P = 0.02$). 20 mM NAC (unpaired *t*-test, $P = 0.03$) and 1 μM PLX4720 (two-tailed unpaired *t*-test, $P = 0.006$) compared with control. **e**, Live–dead staining of spheroids from **d** (ANOVA, $P < 0.0001$; Holm–Šídák multiple comparisons test indicated ($P < 0.0001$) after pre-treatment with 20 mM NAC in presence of either control or 1 μM PLX4720). **f**, β-Catenin staining of biopsies (original magnification ×200) from patients undergoing vemurafenib treatment (percentage indicates percentage RECIST response). **g**, Tabulation of patient samples correlating H-score (intensity of stain per field of cells) to RECIST response (two-tailed paired *t*-test, $P = 0.035$); 30% is considered a responder by RECIST criteria. **h**, Young mice ($n = 10$ per group) with Yumm1.7 tumours were treated with rsFRP2 (200 ng ml^{-1}) twice a week. PLX4720 (417 mg kg^{-1}) was administered once the tumour reached 500 mm^3. rsFRP2-treated (black line) versus control (red line) young mice (ANOVA, $P = 0.009$; Holm–Šídák corrected multiple comparisons, $P < 0.05$ after day 12, mean ± s.e.m.). **i**, Yumm1.7 tumours were injected in aged mice (52 weeks, $n = 5$ per group) pre-treated with α-sFRP2 antibody (1 mg kg^{-1}, once weekly). Mice were administered control or 417 mg kg^{-1} PLX4720-laced chow. Tumour volume at day 25 is shown (unpaired *t*-test with Welch's correction, $P = 0.048$). **j**, RECIST response in patients younger than 65 years versus those older than 65 years. Two-sample Wilcoxon rank-sum (Mann–Whitney) test indicated statistical significance ($P = 0.016$). Data represented as mean ± s.d. (**a**–**c**, **e**, **i**, **j**).

significant difference could be observed. We observed a striking separation ($P = 0.016$) between patients under 65 years and over 65 years (~25.6% reduction in tumour burden in aged compared with ~47% in the young, Fig. 5j). The continuous relationship between response and age was highly significant (Spearman's correlation coefficient of $r = 0.243$, $P = 0.03$) (Extended Data Fig. 9f). However, in this small sample set, the relationship between patient age, sFRP2, and RECIST was not significant. Together, these data suggest that sFRP2-induced β-catenin loss not only promotes invasion, but also renders melanoma cells more sensitive to oxidative stress. New data from ref. 24 show that therapy-refractory, BRAF mutant cells are sensitive to anti-oxidants, specifically vitamin C. Our data support these observations, and suggest that aged patients may uniquely benefit from anti-oxidant therapy.

In our experiments, genetically identical cells had a very different outcome in terms of metastasis and therapy response when placed in an aged microenvironment. In the skin, benign melanocytic lesions (naevi) bear mutational changes that do not result in tumorigenesis, until cells accumulate further changes during ageing. These changes may be genetic, such as loss of p16, but they may also be epigenetic, and age-related epigenetic changes also contribute to tumorigenesis[25–27]. What drives the initiation of these epigenetic changes is not well understood, and it is possible that the aged secretome contributes to this process. While overall results were consistent among different fibroblasts of similar ages, we did observe some phenotypic differences. These may be attributable to factors such as tanning, as, indeed, the exposure of melanoma cells to ultraviolet irradiation not only drives tumour promotion[28] but also metastasis[29]. Our data suggest that, as the general population ages, new efforts must be made to understand and treat cancer in an age-appropriate manner.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 24 November 2014; accepted 2 February 2016.
Published online 4 April 2016.

1. Ruiter, D., Bogenrieder, T., Elder, D. & Herlyn, M. Melanoma-stroma interactions: structural and functional aspects. *Lancet Oncol.* **3,** 35–43 (2002).
2. Li, G., Satyamoorthy, K. & Herlyn, M. Dynamics of cell interactions and communications during melanoma development. *Crit. Rev. Oral Biol. Med.* **13,** 62–70 (2002).
3. Hsu, M. Y., Meier, F. & Herlyn, M. Melanoma development and progression: a conspiracy between tumor and host. *Differentiation* **70,** 522–536 (2002).
4. Bogenrieder, T. & Herlyn, M. Cell-surface proteolysis, growth factor activation and intercellular communication in the progression of melanoma. *Crit. Rev. Oncol. Hematol.* **44,** 1–15 (2002).
5. Dankort, D. *et al.* $Braf^{V600E}$ cooperates with *Pten* loss to induce metastatic melanoma. *Nature Genet.* **41,** 544–552 (2009).
6. Hoek, K. S. MITF: the power and the glory. *Pigment Cell Melanoma Res.* **24,** 262–263 (2011).
7. Hoek, K. S. *et al. In vivo* switching of human melanoma cells between proliferative and invasive states. *Cancer Res.* **68,** 650–656 (2008).
8. Webster, M. R., Kugel, C. H., III & Weeraratna, A. T. The Wnts of change: how Wnts regulate phenotype switching in melanoma. *Biochim. Biophys. Acta* **1856,** 244–251 (2015).
9. O'Connell, M. P. *et al.* Hypoxia induces phenotypic plasticity and therapy resistance in melanoma via the tyrosine kinase receptors ROR1 and ROR2. *Cancer Discov.* **3,** 1378–1393 (2013).
10. Flach, E. H., Rebecca, V. W., Herlyn, M., Smalley, K. S. & Anderson, A. R. Fibroblasts contribute to melanoma tumor growth and drug resistance. *Mol. Pharm.* **8,** 2039–2049 (2011).
11. Campisi, J. The role of cellular senescence in skin aging. *J. Investig. Dermatol. Symp. Proc.* **3,** 1–5 (1998).
12. Coppé, J. P., Desprez, P. Y., Krtolica, A. & Campisi, J. The senescence-associated secretory phenotype: the dark side of tumor suppression. *Annu. Rev. Pathol.* **5,** 99–118 (2010).
13. Park, H. W. Biological aging and social characteristics: gerontology, the Baltimore city hospitals, and the national institutes of health. *J. Hist. Med. Allied Sci.* **68,** 49–86 (2013).
14. Arozarena, I. *et al.* In melanoma, beta-catenin is a suppressor of invasion. *Oncogene* **30,** 4531–4543 (2011).
15. Chien, A. J. *et al.* Activated Wnt/β-catenin signaling in melanoma is associated with decreased proliferation in patient tumors and a murine melanoma model. *Proc. Natl Acad. Sci. USA* **106,** 1193–1198 (2009).
16. Bouchlaka, M. N. *et al.* Aging predisposes to acute inflammatory induced pathology after tumor immunotherapy. *J. Exp. Med.* **210,** 2223–2237 (2013).
17. Lento, W. *et al.* Loss of β-catenin triggers oxidative stress and impairs hematopoietic regeneration. *Genes Dev.* **28,** 995–1004 (2014).
18. Widlund, H. R. *et al.* β-Catenin-induced melanoma growth requires the downstream target *Microphthalmia*-associated transcription factor. *J. Cell Biol.* **158,** 1079–1087 (2002).
19. Liu, F., Fu, Y. & Meyskens, F. L. Jr. MiTF regulates cellular response to reactive oxygen species through transcriptional regulation of APE-1/Ref-1. *J. Invest. Dermatol.* **129,** 422–431 (2009).
20. Corazao-Rozas, P. *et al.* Mitochondrial oxidative stress is the achille's heel of melanoma cells resistant to Braf-mutant inhibitor. *Oncotarget* **4,** 1986–1998 (2013).
21. Yu, L. *et al.* Involvement of superoxide and nitric oxide in BRAF inhibitor PLX4032-induced growth inhibition of melanoma cells. *Integr. Biol.* **350,** 1391–1396 (2014).
22. Biechele, T. L. *et al.* Wnt/β-catenin signaling and AXIN1 regulate apoptosis triggered by inhibition of the mutant kinase $BRAF^{V600E}$ in human melanoma. *Sci. Signal.* **5,** ra3 (2012).
23. Konieczkowski, D. J. *et al.* A melanoma cell state distinction influences sensitivity to MAPK pathway inhibitors. *Cancer Discov.* **4,** 816–827 (2014).
24. Yun, J. *et al.* Vitamin C selectively kills KRAS and BRAF mutant colorectal cancer cells by targeting GAPDH. *Science* **350,** 1391–1396 (2015).
25. Issa, J. P. CpG-island methylation in aging and cancer. *Curr. Top. Microbiol. Immunol.* **249,** 101–118 (2000).
26. Issa, J. P. Aging and epigenetic drift: a vicious cycle. *J. Clin. Invest.* **124,** 24–29 (2014).
27. Shah, P. P. *et al.* Lamin B1 depletion in senescent cells triggers large-scale changes in gene expression and the chromatin landscape. *Genes Dev.* **27,** 1787–1799 (2013).
28. Viros, A. *et al.* Ultraviolet radiation accelerates BRAF-driven melanomagenesis by targeting TP53. *Nature* **511,** 478–482 (2014).
29. Bald, T. *et al.* Ultraviolet-radiation-induced inflammation promotes angiotropism and metastasis in melanoma. *Nature* **507,** 109–113 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank D. Altieri, R. Marais, Z. Ronai, M. McMahon, and B. Vogelstein for comments on the manuscript. We thank R. Somasundaram for advice on immune analyses, M. Herlyn for the WM cell lines, and G. Bollag for PLX4720. We also thank R. Delgiacco, D. Gourevitch, F. Keeney, and D. Schultz. We thank A. Dias-Wanigasekera, E. Gaddy, and M. Ha for technical assistance, and R. Locke for editing the manuscript. This work was supported in part by funds from the Intramural Program of the National Institute on Aging, Baltimore, Maryland (N.M., K.G.B., R.M., W.H.W., L.F.), The Harry J. Lloyd Foundation (K.M., A.T.W.), P01 CA 114046-06 (A.T.W., Q.L.), T32 CA 9171-36 (M.R.W., C.H.K.), an ACS-IRG award (A.T.W.), the Melanoma Research Foundation (A.T.W.), and RO1 CA174746-01 (A.T.W., A.K.). Core facilities at the Wistar are supported by Cancer Center Support Grant P30 CA010815.

Author Contributions A.T.W. conceived and designed the project. A.T.W. and A.K. designed and supervised the experiments. A.K., M.R.W., K.M., R.B., A.N., C.H.K., V.M.D., J.A., M.P.O., P.C., A.A.V., W.H.W., E.L., and K.M.A. performed the experiments. A.T.W., A.K., A.V.K., H.Y.T., X.Y., E.L., Z.E., K.G.B., R.Z., X.X., Q.L., and D.W.S. analysed the experimental data. A.T.W., A.K., and Q.L. designed and supervised data analysis and statistical analysis. M.B., A.R., D.S., J.S., B.S., R.S.L., M.C., A.M.M., G.V.L., D.B.J., R.M., N.B.M., L.F., K.M., K.T.F., D.T.F., J.A.W., Z.A.C., M.T.T., C.H., E.B., and G.K. performed data collection and provided anonymized patient data and samples and reagents. A.T.W. and A.K. wrote the manuscript. All authors discussed the results and commented on the manuscript.

Author Information Microarray data are available in the GEO database under accession number GSE57445. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.T.W. (aweeraratna@wistar.org).

METHODS

The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Cell culture. FS5, FS4, FS13, FS14, M93-047, UACC-903, and UACC-1273 cells were maintained in RPMI (Invitrogen), supplemented with 10% FBS, 100 units per millilitre penicillin and streptomycin, and 4 mM L-glutamine. WM35, WM793, WM164, WM1799, and 1205 LU cells were maintained in MCDB153 (Sigma)/L-15 (Cellgro) (4:1 ratio) supplemented with 2% FBS and 1.6 mM $CaCl_2$ (tumour growth media). WM983b and WM3918 cells were maintained in DMEM (Invitrogen), supplemented with 5% FBS, 100 units per millilitre penicillin and streptomycin, and 4 mM L-glutamine. YUMM1.7 cells were maintained in DMEM F-12 (HEPES/glutamine) supplemented with 10% FBS, 1% NEAA, and 100 units per millilitre penicillin and streptomycin. Fibroblasts were maintained in DMEM, supplemented with 10% FBS, 100 units per millilitre penicillin and streptomycin, and 4 mM L-glutamine. Keratinocytes were maintained in keratinocyte SFM supplemented with human recombinant Epidermal Growth Factor 1-53 (EGF 1-53) and Bovine Pituitary Extract (BPE) (Invitrogen). Cell lines were cultured at 37 °C in 5% CO_2 and the medium was replaced as required. Cell stocks were fingerprinted using an AmpFLSTR Identifiler PCR Amplification Kit from Life Technologies at The Wistar Institute Genomics Facility. Although it is desirable to compare the profile with the tissue or patient of origin, our cell lines were established over the course of 40 years, long before acquisition of normal control DNA was routinely performed. However, each short tandem repeat profile is compared with our internal database of over 200 melanoma cell lines, as well as control lines, such as HeLa and 293 T. Short tandem repeat profiles are available upon request. Cell culture supernatants were tested for mycoplasma using a Lonza MycoAlert assay at the University of Pennsylvania Cell Center Services.

Organotypic three-dimensional skin reconstructs. Organotypic three-dimensional skin reconstructs were generated as previously described[30]. In each insert, 6.4×10^4 fibroblasts were plated on top of the acellular layer (BD 355467 and Falcon 353092) and incubated for 45 min at 37 °C in a 5% CO_2 tissue culture incubator. DMEM containing 10% FBS was added to each well of the tissue culture trays and incubated for 4 days. Reconstructs were then incubated for 1 h at 37 °C in HBSS containing 1% dialysed FBS (wash media). Washing media were removed and replaced with reconstruct media I. Keratinocytes (4.17×10^5) and melanoma cells (8.3×10^4) were added to the inside of each insert. Media were changed every other day until day 18 when reconstructs were harvested, fixed in 10% formalin, paraffin embedded, sectioned, and stained. Quantification of the invasion was performed using ImageJ software (available at http://imagej.nih.gov/ij/; developed by W. Rasband).

Three-dimensional spheroid assays. Tissue-culture-treated 96-well plates were coated with 50 μl 1.5% Difco Agar Noble (Becton Dickinson). Melanoma cells were seeded at 5×10^3 cells per well and allowed to form spheroids over 72 h. Spheroids were harvested and embedded as previously described using collagen type I (GIBCO, A1048301). For spheroids incubated with fibroblast conditioned media, fibroblasts were seeded onto 75 cm^2 flasks at 7×10^5 to 9×10^5 per flask depending on growth rate. Sixteen hours later, media were replaced and incubated for 72 h. Media from young fibroblasts were combined and media from aged fibroblasts were combined. These conditioned media were added to the top of the collagen plug containing the spheroids. Quantitation of invasive surface area was performed using NIS Elements Advanced Research software.

Live–dead staining. Spheroids were generated and embedded as described above. Spheroids were stained using a LIVE/DEAD Viability/Cytotoxicity Kit (L3224, Invitrogen). Briefly, spheroids were washed with PBS and stained with calcein AM/Ethidium homodimer-1. The dyes were diluted in PBS and 300 μl of the solution was added on the spheroid wells for 1 h at 37 °C. The spheroids were washed in PBS and imaged using a Nikon TE2000 Inverted Microscope. Quantitation of fluorescence intensity was performed using NIS Elements Advanced Research software.

Boyden chamber invasion assays. Matrigel (BD Biosciences, 354234) was diluted in PBS (1:3,000 dilution). One hundred and fifty microlitres of this mixture was pipetted into each insert of the invasion assay plate (Corning, 3422). The plate was incubated at 37 °C for 2 h and then dried at room temperature (25 °C) overnight under sterile conditions. Melanoma cells were pre-treated for 48 h in six-well plates. After 48 h, the cells were harvested and 1.5×10^5 cells were added to each transwell. High concentration serum media (RPMI with 20% FCS, tumour growth media with 10% FCS) were added to the outside (bottom) of the well. The plates were incubated at 37 °C until cells had migrated to the bottom of the well. The migrated cells were fixed in 95% ice-cold methanol and stained with crystal violet (0.5%) for 10 min. The stain was washed and the wells were left to dry. Cells were imaged and quantified using ImageJ software.

Cellular proliferation assays. In a 24-well plate, 5,000 cells in triplicate were plated per day of measurement. Every 2–3 days, cells were counted using a haemocytometer and the total cell number in the well was recorded and plotted on GraphPad Prism.

Immunofluorescence. Cells were seeded onto glass cover slips at 1×10^4 to 4×10^4 cells per well, and incubated overnight. After treatment, cells were fixed using 95% methanol. Primary antibodies were diluted as stated above in blocking buffer and incubated overnight at 4 °C. Cells were washed in PBS and incubated with the appropriate secondary antibody (1:2,000, Invitrogen) for 1 h at room temperature. Cells were then washed in PBS and mounted in Prolong Gold anti-fade reagent containing DAPI (Invitrogen). Images were captured on a Leica TCS SP5 II scanning laser confocal system.

Immunohistochemistry. All antibodies are described in the Supplementary Information. Patient samples were collected under IRB exemption approval for protocol EX21205258-1. Paraffin embedded sections were rehydrated through a xylene and alcohol series, rinsed in H_2O and washed in PBS. Antigen retrieval was performed using target retrieval buffer (Vector Labs) and steamed for 20 min. Samples were then blocked in a peroxidase blocking buffer (Thermo Scientific) for 15 min, followed by Protein block (Thermo Scientific) for 5 min, and incubated in appropriate primary antibody diluted in antibody diluent (S0809, Dako) at 4 °C overnight in a humidified chamber. For mouse samples to be incubated with anti-mouse antibody, samples were blocked for 1 h in Mouse on Mouse (M.O.M.) Blocking Reagent (MKB-2213, Vector Labs). After washing in PBS, samples were incubated in biotinylated anti-rabbit or polyvalent secondary antibody (Thermo Scientific) followed by streptavidin-HRP solution at room temperature for 20 min. Samples were then washed in PBS and incubated in 3-amino-9-ethyl-l-carboazole (AEC) chromogen and counterstained with Mayer's haematoxylin for 1 min, rinsed in cold H_2O, and mounted in Aquamount.

Western blotting. Total protein lysate (50–65 μg) was run on a 4–12% NuPAGE Bis Tris gel (Invitrogen). Proteins were then transferred onto PVDF membrane using an iBlot system, and blocked in 5% milk/TBST for 1 h. All primary antibodies were diluted in 5% milk/TBST and incubated over night at 4 °C. The membranes were washed in TBST and probed with the corresponding HRP-conjugated secondary antibody (0.2–0.02 μg ml^{-1} of anti-mouse, streptavidin, or anti-rabbit). Proteins were visualized using ECL prime (Amersham), or Luminata Crescendo (Millipore).

Lentiviral infection. All clones used are described in the Supplementary Information. All short hairpin RNA (shRNA) was obtained from the TRC shRNA library available through the Molecular Screening Facility at The Wistar Institute. Lentiviral production was performed as described in the protocol developed by the TRC library (Broad Institute). Briefly, 293 T cells were co-transfected with shRNA vector and lentiviral packaging plasmids (pCMV-dR8.74psPAX2, pMD2.G). The supernatant containing virus was harvested at 36 and 60 h, combined and filtered through a 0.45 μm filter. For transduction, the cells were layered overnight with lentivirus containing 8 μg ml^{-1} polybrene. The cells were allowed to recover for 24 h and then selected using 1 μg ml^{-1} puromycin.

MTS assays. Cells (1.5×10^3 per well) were plated in a 96-well plate and treated with PLX4720. After 48 h, cells were incubated with MTS dye (20 μl per well) for 2 h. Absorbance was determined at 490 nm using an EL800 microplate reader (BioTek). The percentage cell proliferation was calculated by converting the experimental absorbance to percentage of control and plotted versus drug concentration. The values were then analysed using a nonlinear dose–response analysis in GraphPad Prism.

Illumina oligonucleotide microarray. Transcriptional profiling was determined using Illumina Sentrix BeadChips. Total RNA was used to generate biotin-labelled cRNA using the Illumina TotalPrep RNA Amplification Kit. In short, 0.5 μg of total RNA was first converted into single-stranded complementary DNA (cDNA) with reverse transcriptase using an oligo-dT primer containing the T7 RNA polymerase promoter site and then copied to produce double-stranded cDNA molecules. The double-stranded cDNA was cleaned and concentrated with the supplied columns and used in an overnight *in vitro* transcription reaction where single-stranded RNA (cRNA) was generated incorporating biotin-16-UTP. A total of 0.75 μg of biotin-labelled cRNA was hybridized at 58 °C for 16 h to Illumina's Sentrix Human HT-12 v3 Expression BeadChips (Illumina). Each BeadChip has around 48,000 transcripts with approximately 15-fold redundancy. The arrays were washed, blocked, and the labelled cRNA was detected by staining with streptavidin-Cy3. Hybridized arrays were scanned using an Illumina BeadStation 500X Genetic Analysis Systems scanner and the image data extracted using the Illumina GenomeStudio software, version 1.1.1). Data are available in the GEO database (accession number GSE57445).

Microarray analysis. Microarray expression data were quantile normalized and probes that showed low expression levels (detection P value > 0.05) across all samples were removed from the analysis. Expression values for each cell line were tested separately in multiple linear regression model with fibroblast age and experiment batches as predictor variables. Matlab version 8.0 'regress' function was used to calculated P values for each probe for association with fibroblast age.

False discovery rate was estimated using Benjamini-Hochberg procedure; only probes that showed a false discovery rate $< 5\%$ in all three cell lines were considered significant. Heat map was plotted using average expression values for three groups of age (young, middle, and aged) normalized to aged group (100%).

***In vivo* tail vein metastases assay.** All animal experiments were approved by the Institutional Animal Care and Use Committee (112503Y_0) and were performed in a facility accredited by the Association for the Assessment and Accreditation of Laboratory Animal Care. From preliminary studies, we observed significant differences (more than 1.8 standard deviations, a very large effect size) in some of the outcomes between young and aged groups. As few as five samples in each group in this study afforded 80% power at a two-sided α of 0.05 to detect a difference of about 1.8 standard deviations in a continuous outcome between young and aged groups, but we increased the sample size slightly to account for potential loss of mice due to health issues associated with ageing. Male C57BL6 mice at 6–8 weeks (young) and 52 weeks (aged) were purchased from Taconic. YUMM 1.7 (1×10^6 cells per 100 µl PBS) or B16F10 (2.5×10^5 per 100 µl PBS) were injected into the tail vein of C57BL6 mice. Alternatively, Yumm1.7 cells were overexpressed with mCherry plasmid (pLU-EF1-MCS-mCherry) using lentivirus. The cells were sorted for mCherry and 1×10^6 cells per 100 µl PBS were injected into tail vein of young C57BL6 mice. After 4 weeks, the mice were euthanized, lungs were harvested, and metastases counted. Lungs were fixed in paraffin and stained with haematoxylin and eosin. Alternatively, lungs were harvested and imaged for presence of metastatic melanoma cells using a Perkin-Elmer IVIS 200 whole body imager. For experiments requiring rsFRP2, mouse rsFRP2 (1169-FR-025/CF, R&D) was diluted in 50 µl PBS and injected at a concentration of 200 ng per mouse twice a week. The levels of sFRP2 were monitored by submandibular blood withdrawal every 2 weeks.

***In vivo* PLX4720 assay.** All animal experiments were approved by the Institutional Animal Care and Use Committee (112503X_0) and were performed in a facility accredited by the Association for the Assessment and Accreditation of Laboratory Animal Care. From preliminary studies, we observed significant differences (more than 1.8 standard deviations, a very large effect size) in some of the outcomes between young and aged groups. As few as five samples in each group in this study afforded 80% power at a two-sided α of 0.05 to detect a difference of about 1.8 standard deviations in a continuous outcome between young and aged groups, but we increased the sample size slightly to account for potential loss of mice due to health issues associated with ageing. YUMM1.7 (2.5×10^5 cells) were suspended in Matrigel (500 µg ml−1) and injected subcutaneously into young (6 week) and aged (52 week) C57/BL6 mice (Taconic). When resulting tumours reached 200 mm3, mice were fed either AIN-76A chow or AIN-76A chow containing 417 mg kg^{-1} PLX4720. Tumour sizes were measured every 3–4 days using digital callipers, and tumour volumes were calculated using the following formula: volume $= 0.5 \times$ (length $\times$ width^2). Time-to-event (survival) was determined by a fivefold increase in baseline volume ($\sim$1,000 mm^3) and was limited by the development of skin necrosis. Upon the occurrence of necrosis, mice were euthanized. For subsequent experiments involving sFRP2 manipulation, Yumm1.7 cells overexpressing mCherry were used. One million cells were suspended in PBS and subcutaneously injected into either 6-week-old or 52-week-old male C57/BL6 mice (Taconic). For treatment with rsFRP2, the mice were injected with recombinant protein (200 ng ml^{-1}) through the tail vein every 2 days as described above. For the experiments performed with sFRP2 blocking antibody (clone 80.8.6, MABC539, EMD Millipore), the mice were treated with 1 mg kg^{-1} antibody (either sFRP2 or isotype control, Biolegend, 400264) once a week through tail vein injections.

Comet assay. WM35 and FS5 melanoma cells were seeded in 12-well plates and treated with conditioned media for 48 h. The cells were harvested in ice-cold PBS. The comet assay was performed using CometSlides (Trevigen). Briefly, 75 µl of a 2×10^5 cell suspension was mixed with 500 µl 1% low melting point agarose. Fifty microlitres of cell/agarose mixture was dropped into the wells and allowed to solidify. Slides were incubated in lysis buffer (1.2 M NaCl, 100 nM EDTA, 0.1% Sarkosyl, pH 10.0) for 1 h at 4 °C. Slides were then electrophoresed at 25 V for 12 min in alkaline buffer (0.03 M NaOH, 2 mM EDTA, pH 8.0). After fixation in 70% ethanol, comets were visualized by staining with SYBR Green (Fisher). The extent of DNA damage was measured as the artificial Olive Moment using Cometscore software downloaded from http://www.tritekcorp.com.

ROS activation assay. Five thousand melanoma cells were seeded in a 96-well plate in triplicate. The cells were then treated with conditioned media from young and aged fibroblasts as well as DMEM as control. Hydrogen peroxide was added at 1 mM as control. After 72 h, ROS were measured using a Cell Meter Fluorimetric Intracellular Total ROS Activity Assay Kit (22901, AAT Bioquest) according to the manufacturer's protocol. The plates were measured using a PerkinElmer EnVision Xcite Multilabel plate reader using the filters for excitation/emission (Ex/Em) = 520/605 nm. Alternatively, samples were imaged using PerkinElmer Operetta and the fluorescent signal was quantified with Harmony 3.0 software. The cell number was determined by Hoescht staining (Hoechst 33342, Invitrogen) and used to normalize the total fluorescence obtained from ROS staining.

Topflash assay. Topflash vectors were obtained from Addgene (M51 Super 8x FOPFlash/TOPFlash mutant, 12457; M50 Super 8x TOPFlash, 12456). WM35 cells were plated to achieve 70% confluency in six-well plates. Cells were co-transfected with pTK-RLuc (Green Renilla Luciferase) along with either Topflash or Fopflash vectors. After 5 h of transfection, cells were treated as required. After 48 h, cells were harvested and luciferase activity was measured using a Dual-Luciferase Reporter (DLR) Assay System (Promega, E1910). The firefly luciferase signal from each well was normalized to its *Renilla* luciferase signal. Topflash/fopflash signal was determined from each treatment and graphed using Graphpad/Prism.

Treatment with NAC. NAC was obtained from Sigma (A9165) and dissolved in sterile distilled H_2O (stock 1 M). Cells were treated for 48 h and analysed. After optimization, 10 mM final concentration was used for subsequent experiments.

Cell fractionation. Melanoma cells were seeded into T25 flasks and incubated for 72 h with 6.5 ml conditioned fibroblast media prepared as described above. Cells were then washed in PBS, harvested with TrypLE Express and fractionated using the cellular fractionation kit (NE-PER, Fisher) as per the manufacturer's protocol. Cell lysates were then separated on an SDS–polyacrylamide gel electrophoresis gel and visualized using standard western blotting procedures.

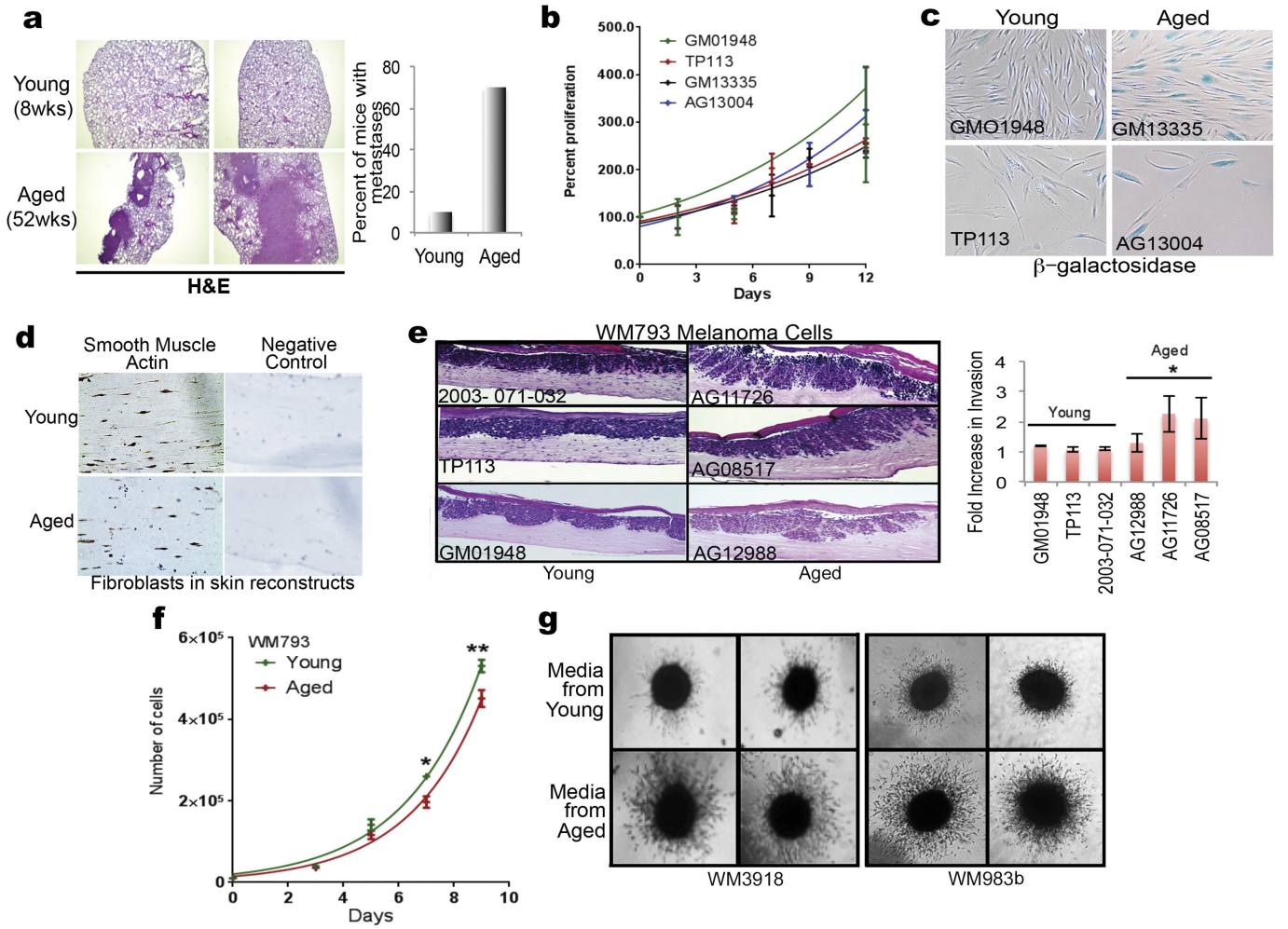
ELISA. Nunc MaxiSorp ELISA plates (ebiosciences) were coated with 50 µl of 3 µg ml^{-1} sFRP2 (ab137560, Abcam) overnight at 4 °C. Plates were washed in PBS containing 0.1% Tween and blocked in ELISA diluent (00-4202-56, eBioscience) for 2 h. Serum was diluted 1:100 before addition to the plates and incubated overnight at 4 °C. The next day, the plates were washed in PBS containing 0.1% Tween20 and incubated with detection antibody (MAB6838, R&D Systems) for 1 h at room temperature. Plates were washed and incubated with secondary antibody for 1 h. After washing, 100 µl TMB (00-4201-56, eBioscience) was added to the plates and incubated for 15 min The reaction was stopped using 50 µl of 2 N H_2SO_4 and absorbance was measured at 450 nm.

β-Galactosidase staining. Fibroblasts were plated into 12-well dishes, incubated for 48 h, washed with PBS, and fixed in 2% formaldehyde/0.2% glutaraldehyde. Cells were then incubated in staining solution (150 mM NaCl, Sigma), 2 mM $MgCl_2$ (Sigma), 5 mM $K_3Fe(CN)_6$ (Millipore), 5 mM $K_4Fe(CN)_6$ (Millipore), 40 mM Na_2PO_4 (Sigma) pH 5.5, 20 mg ml^{-1} X-gal (Applichem, Darmstadt, Germany) at 37 °C overnight. Stain was removed and cells were stored in 70% glycerol before being imaged.

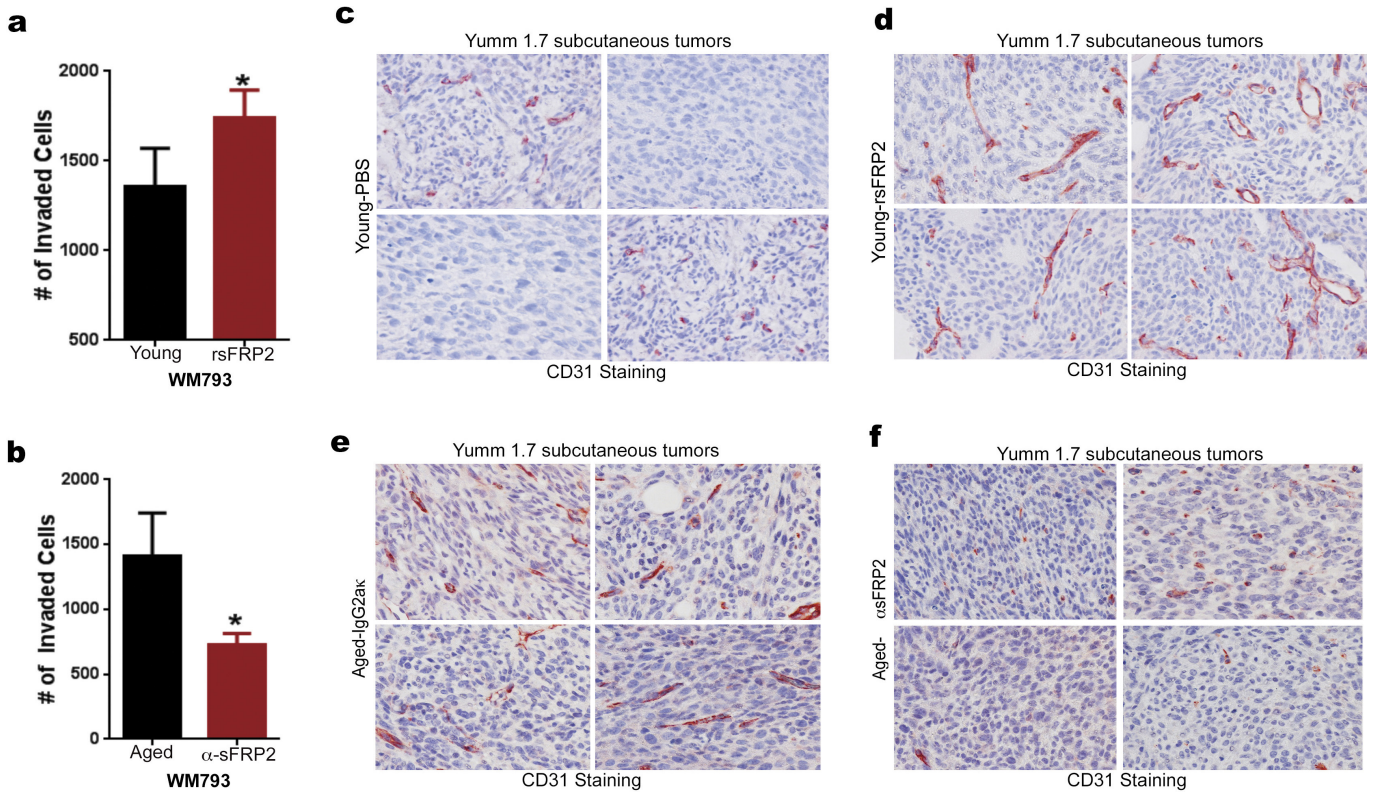
Quantitative PCR. All primers are listed in the Supplementary Information. Mouse tissue was snap frozen in liquid nitrogen immediately after harvesting. Ten milligrams of the lung tissue was homogenized and RNA was extracted using Trizol (Invitrogen) and RNeasy Mini kit (Qiagen) as described previously. One microgram of RNA was used to prepare cDNA using iscript DNA synthesis kit (1708891, Bio-Rad). cDNA was diluted 1:5 before use in further reactions. Each 20 µl well reaction comprised 10 µl Power SYBR Green Master mix (4367659, Invitrogen), 1 µl primer mix (Final concentration 0.5 µM), and 1 µl cDNA. Standard curves were generated for all primers and each set of primers was normalized to an 18 s primer pair.

Statistical analysis. For *in vitro* studies, a Student's *t*-test or Wilcoxon rank-sum (Mann–Whitney) test was performed for two-group comparisons. Estimate of variance was performed and parameters for the *t*-test were adjusted accordingly using Welch's correction. An ANOVA or Kruskal–Wallis test with post-hoc Bonferroni's or Holm–Šídák's adjusted *P* values was used for multiple comparisons. For dose–response analysis, Spearman's correlation was calculated. For *in vivo* studies, the indicated sample size for each experiment was designed to have 80% power at a two-sided α of 0.05 to detect a difference of large effect size of about 1.25 between two groups on a continuous measurement. The fold change in tumour volume at each time point after treatment relative to baseline was calculated and then the fold change in treatment group relative to the age-matched control group was used with a mixed-effect model to evaluate the treatment effect between age groups. Stata 12.0 (StataCorp) was used for data analysis for *in vivo* studies and human samples. For other experiments, Graphpad/Prism6 was used for plotting graphs and statistical analysis. Significance was designated as follows: $*P < 0.05$; $**P < 0.01$; $***P < 0.001$. Extended statistical analyses for patient data are provided in the Supplementary Information.

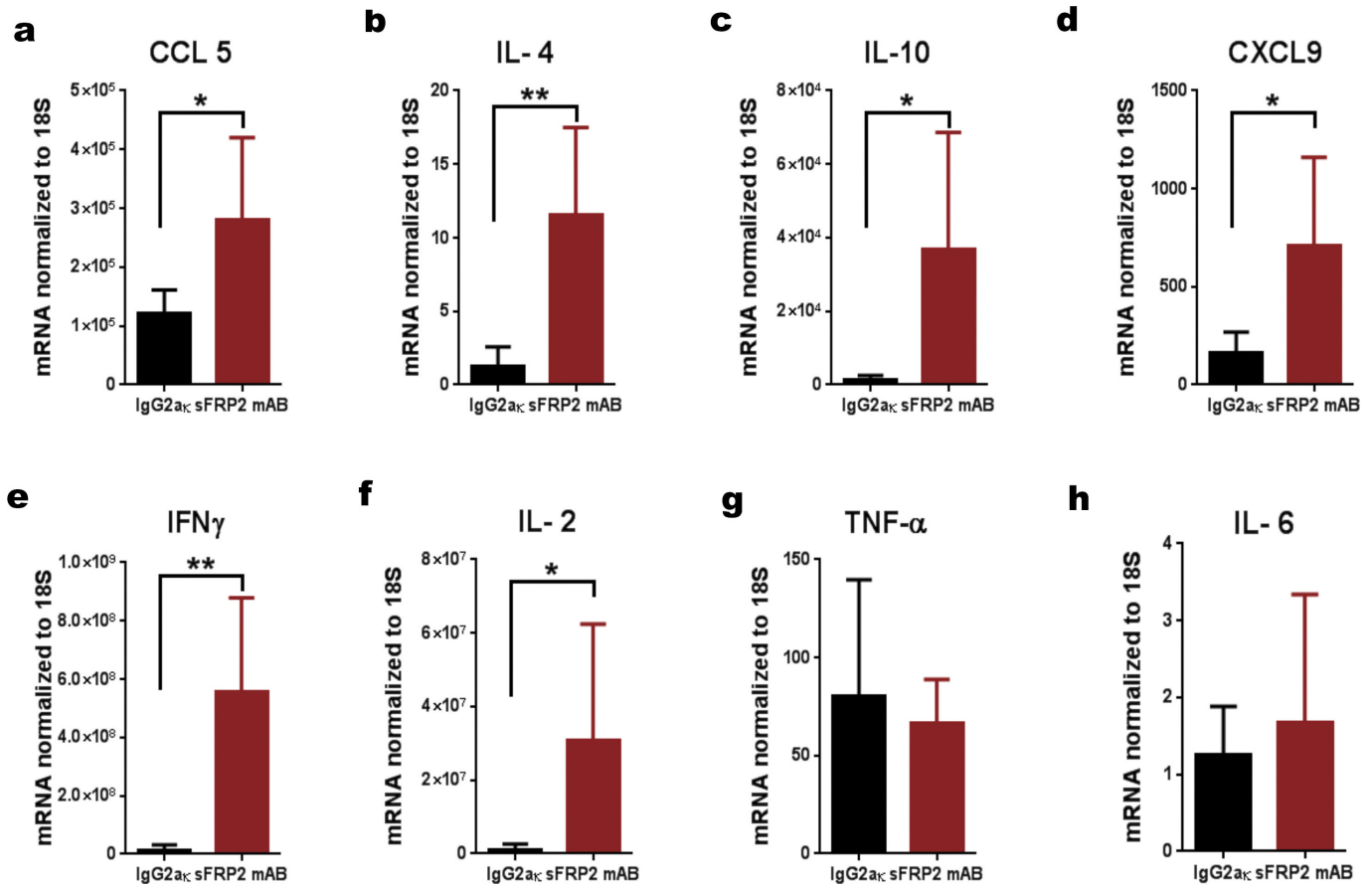
30. Berking, C. & Herlyn, M. Human skin reconstruct models: a new application for studies of melanocyte and melanoma biology. *Histol. Histopathol.* **16,** 669–674 (2001).



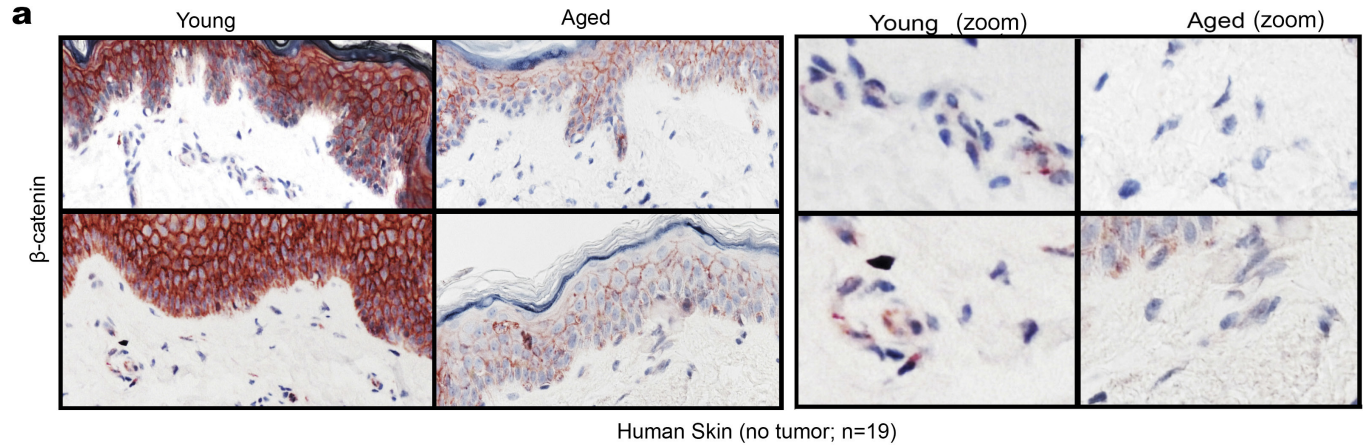
Extended Data Figure 1 | Characterization of young and aged fibroblasts. a, One million Yumm1.7 cells were injected into the tail vein of young (8 weeks, $n = 10$ mice) and aged (52 weeks, $n = 10$ mice) mice; 3 weeks later, lungs were assessed for metastatic colonies. Samples were analysed by haematoxylin and eosin staining. Number of mice with metastatic colonies in the lungs is quantified in the graph. **b**, Proliferation rate of aged and young fibroblasts was measured by simple cell counts over a period of 12 days. ANOVA is insignificant ($P = 0.234$). **c**, Young and aged fibroblasts were assessed for basal β-galactosidase activity after five passages in culture. Representative images from two cell lines are shown for young and aged fibroblasts, original magnification ×100. **d**, Staining of fibroblasts in skin reconstructs with α-SMA-1 to assess persistence of fibroblasts in cell culture. Representative images, original magnification ×150. **e**, WM793 melanoma cells were grown in organotypic three-dimensional skin reconstructs built with three different fibroblast cell lines derived from healthy young (25–35 years) and healthy aged (55–65 years) individuals. Representative images, original magnification ×150. Invasion was quantified using NIS Element software. ANOVA was performed ($P = 0.007$). Holm–Šídák multiple comparisons comparing each young cell line with each aged cell line indicated $P < 0.05$. **f**, WM793 melanoma cells exposed to conditioned media from young and aged fibroblasts were assessed for proliferation using simple cell counts. Repeated measures ANOVA was calculated between samples ($P = 0.006$). Bonferroni's multiple comparisons test on days 7 and 9 was performed to obtain adjusted P value (day 7 ($P = 0.047$), day 9 ($P = 0.0004$)). **g**, Multiple melanoma cells were allowed to form spheroids followed by treatment with conditioned media from aged or young fibroblasts for 48 h. The spheroids were then examined for their ability to invade in a collagen matrix. Data are represented as mean ± s.d. for each graph (**b**, **e**, **f**).

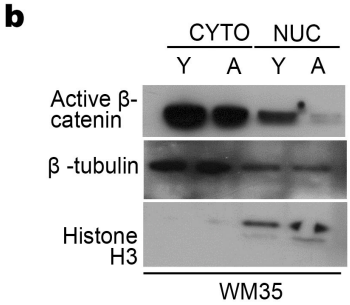


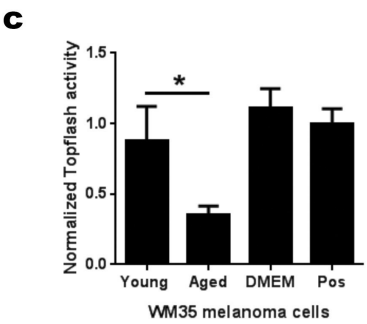
Extended Data Figure 2 | sFRP2 promotes invasion and angiogenesis. **a**, Conditioned media from young fibroblasts treated with either control (PBS) or rsFRP2 (200 ng ml^{-1}) were used to pre-treat WM793 melanoma cells for 48 h. Invasion was assayed using a Boyden chamber assay over 24–72 h. Two-tailed unpaired *t*-test was performed ($P = 0.033$). **b**, Conditioned media from aged fibroblasts treated with sFRP2 blocking antibody (15 μg ml^{-1}) for 72 h were used to pre-treat WM793 melanoma cells for 48 h. The invasion of melanoma cells was assessed in a Boyden chamber assay for 24–72 h. Two-tailed unpaired *t*-test was performed ($P = 0.035$). **c**, **d**, Young mice (8 weeks, 10 per group) were injected subcutaneously with Yumm1.7 cells. After palpable tumour appeared, mice were treated with rsFRP2 (200 ng ml^{-1}) for 30 days and examined for angiogenesis using CD31 staining. Representative images, original magnification ×400. **e**, **f**, Aged mice (52 weeks, $n = 5$ per group) were injected subcutaneously with Yumm1.7 cells and treated with either control (**e**) IgG2aκ or (**f**) sFRP2 blocking antibody (1 mg kg^{-1}) for 3 weeks. Tumours were examined for angiogenesis by CD31 staining. Representative images, original magnification ×400. Data are represented as mean ± s.d. for each graph (**a**, **b**).

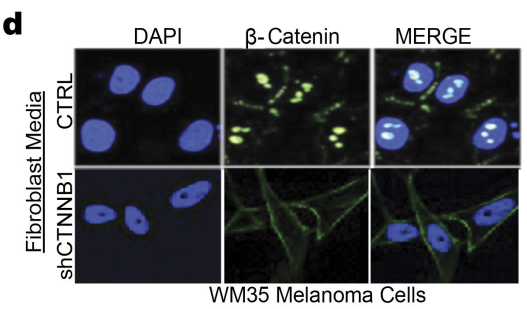


Extended Data Figure 3 | Treatment of aged tumour-bearing mice with an α-sfrp2 antibody results in a lethal inflammation. Cytokine analysis of lungs in aged tumour-bearing mice (52 weeks, $n = 5$ per group) treated with IgG2aκ or α-sFRP2 antibody (1 mg kg^{-1}, once a week for 3 weeks). RT–PCR demonstrates a difference in the lungs of mice treated with IgG2aκ or α-sFRP2 in cytokines (**a**) CCL5, (**b**) IL4, (**c**) IL10, (**d**) CXCL9, (**e**) IFNγ, (**f**) and IL2. Early response inflammatory genes (**g**) *TNF*α and (**h**) *IL6* were no longer significantly altered. Estimate of variance was performed for all genes. For all cytokines, an two-tailed unpaired *t*-test was performed; $*P < 0.05$, $**P < 0.02$. Data are represented as mean ± s.d. for each graph.

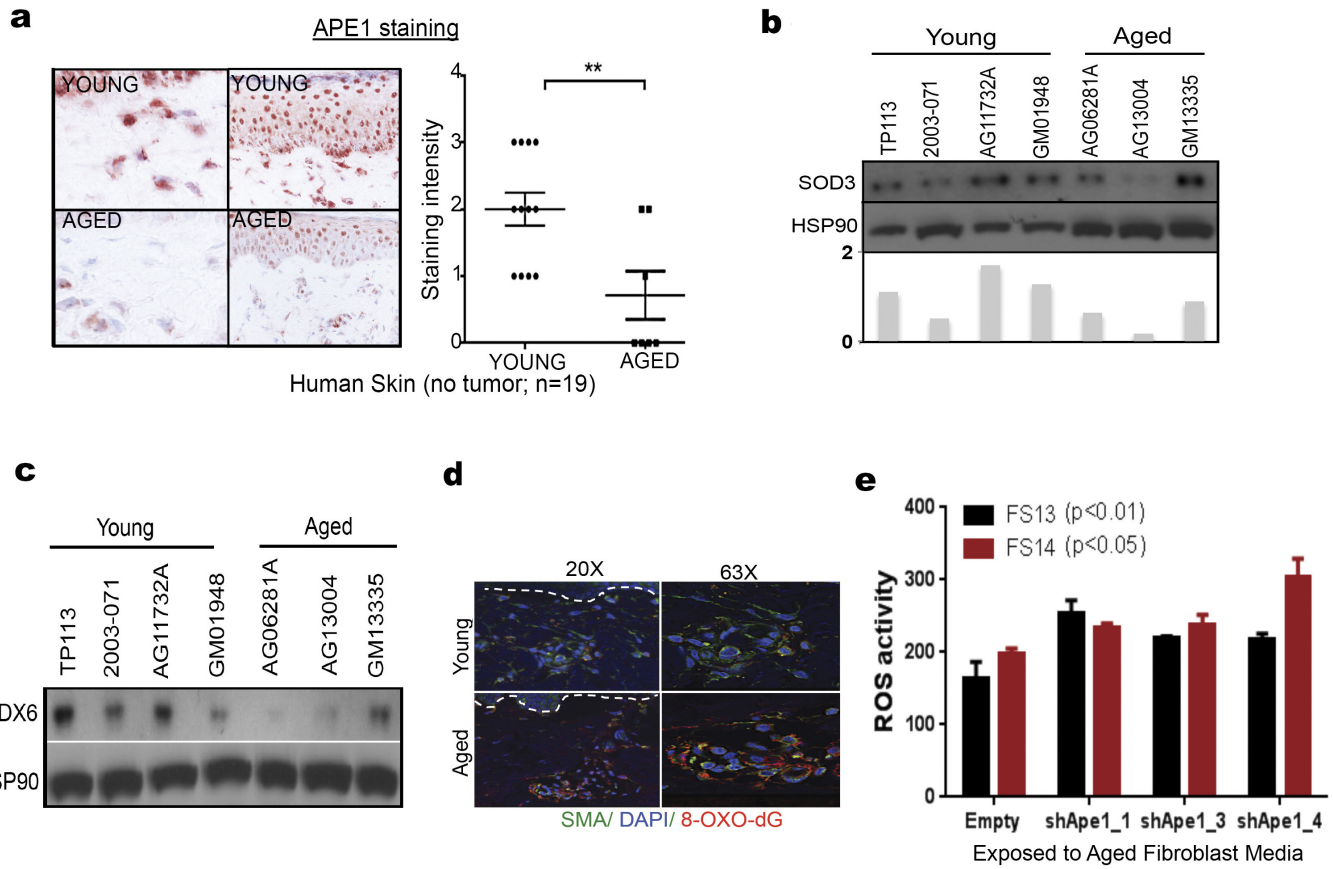




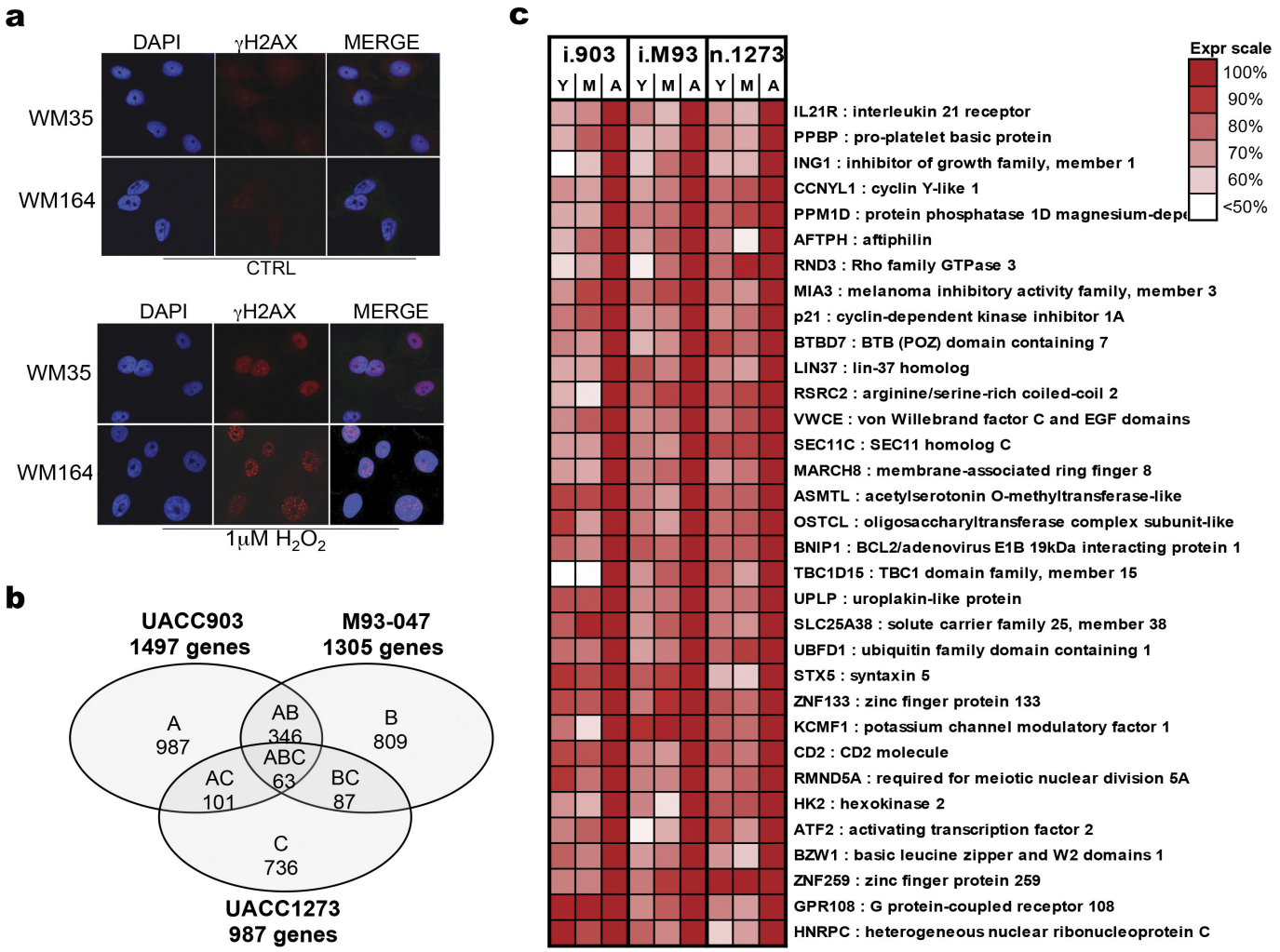




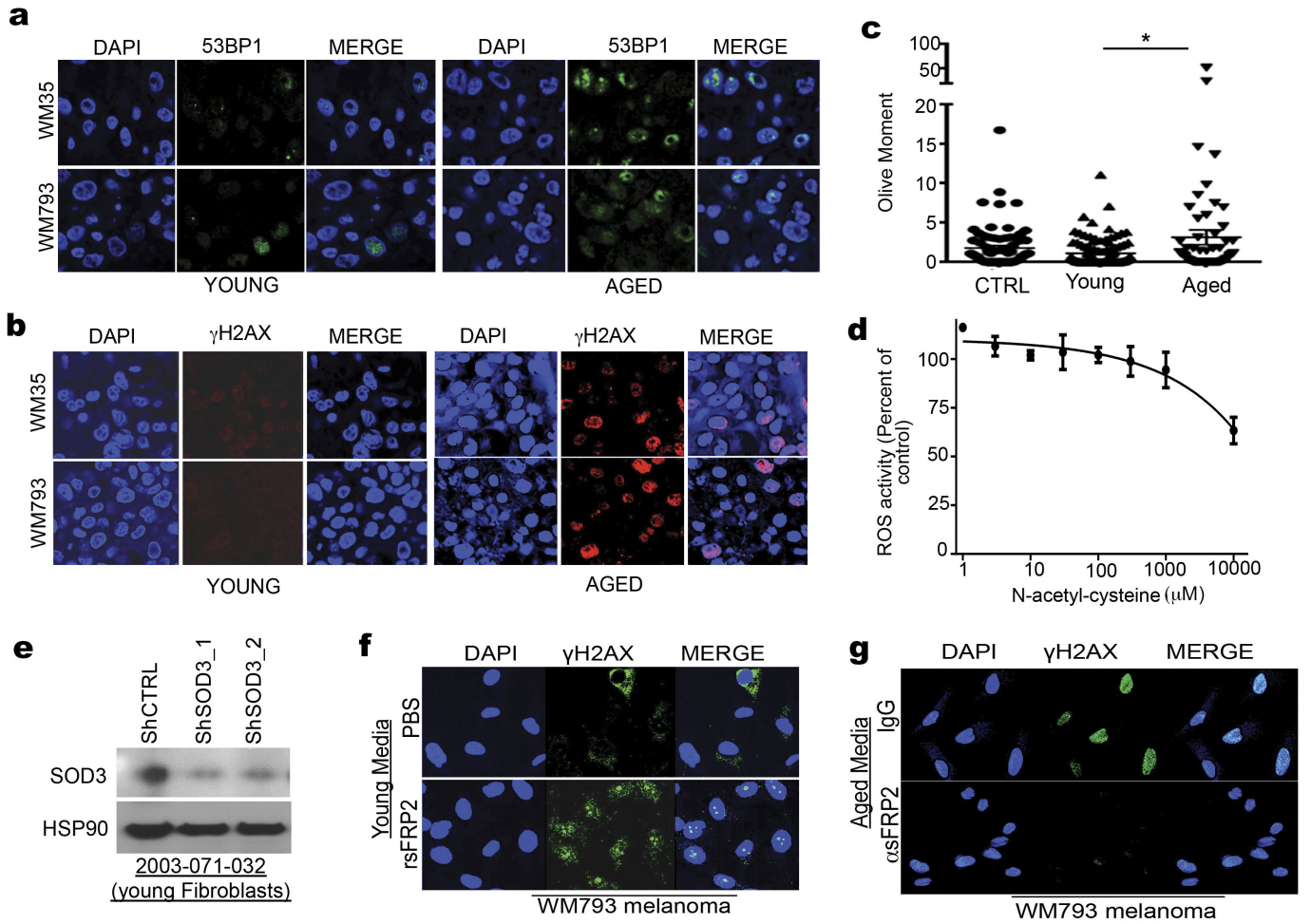
Extended Data Figure 4 | β-Catenin loss in the aged microenvironment. **a**, β-Catenin expression in normal human skin from young and aged donors, with a focus on the fibroblast population (zoom). β-Catenin nuclear translocation in melanoma cells treated with conditioned media from aged compared with young fibroblasts as measured by (**b**) Western analysis and (**c**) a TOPFLASH assay. Two-tailed unpaired *t*-test was performed to indicate statistical significance between treatment with young and aged media ($P = 0.023$). Data are represented as mean ± s.d. (**c**). **d**, Immunofluorescent analysis of β-catenin in melanoma cells treated with media from young fibroblasts in which the β-catenin is knocked down.



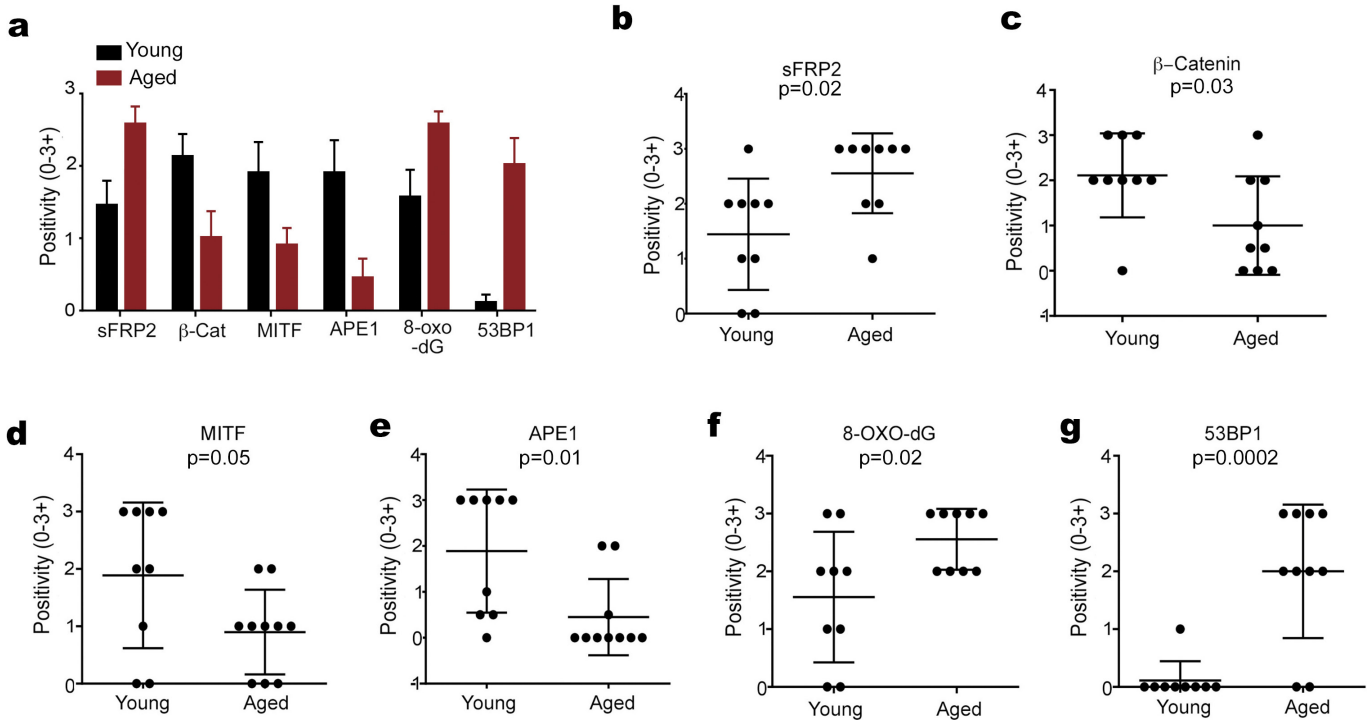
Extended Data Figure 5 | Increase in oxidative stress in the aged microenvironment. a, APE1 expression in normal human skin as measured by immunohistochemistry. Slides were scored for intensity of stain (3, highest; 0, lowest; <35 years, $n = 12$; >55 years, $n = 7$). Representative images, original magnification ×400 (left) and ×200 (right). Unpaired *t*-test using rank sum (Mann–Whitney) revealed statistical significance ($P = 0.009$). Western analysis of (**b**) SOD3 and (**c**) PRDX6 levels in conditioned media from young and aged fibroblasts. **d**, Immunofluorescent analysis of 8-oxo-dG in normal young and aged skin stained for oxidative stress marker 8-oxo-dG (red), smooth muscle actin (green), and DAPI (blue). **e**, ROS activity in melanoma cells with APE1 knockdown, after exposure to aged media. ANOVA was performed for each cell line treatment (FS13 ($P = 0.0006$); FS14 ($P = 0.004$)). For FS13, a two-tailed unpaired *t*-test indicated significance ($P < 0.01$) for each shAPE1 cell line compared with control cells. For FS14, a two-tailed unpaired *t*-test indicated significance ($P < 0.05$) for each shAPE1 cell line compared with control cells. Data are represented as mean ± s.d. for each graph (**a**, **e**).



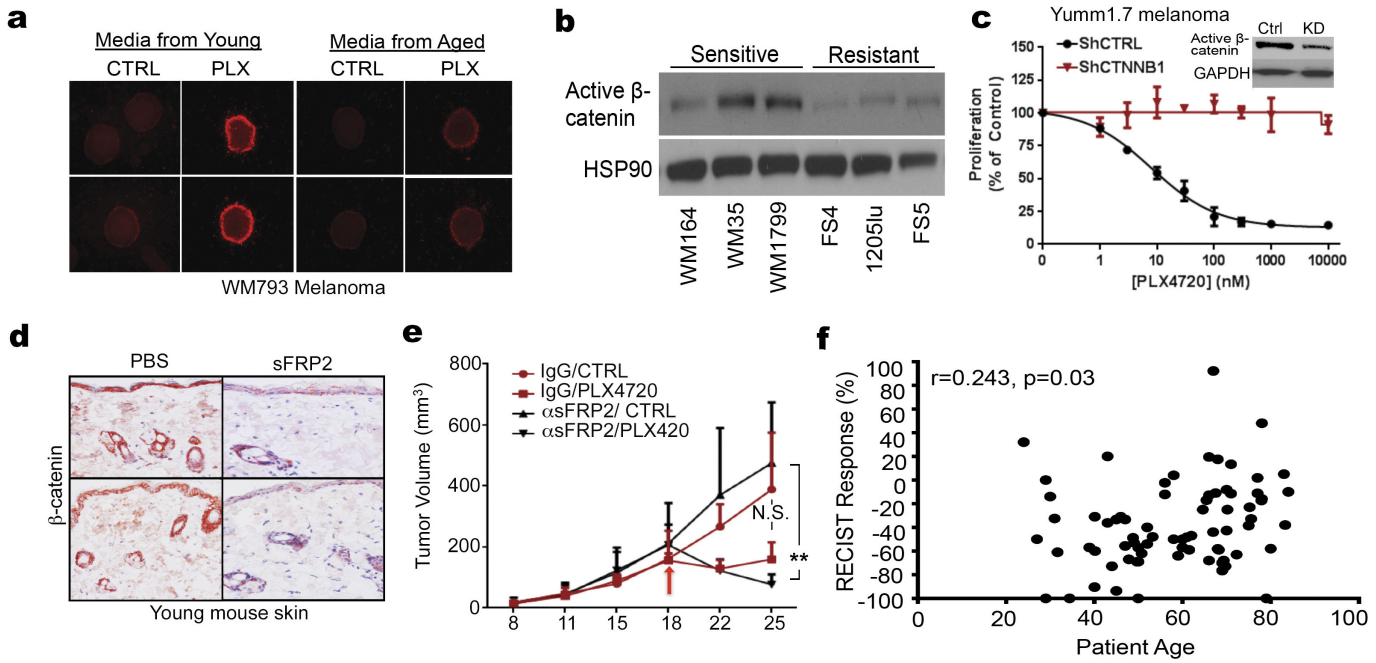
Extended Data Figure 6 | Gene expression analysis of melanoma cells exposed to aged fibroblasts reveals increases in DNA damage. **a**, γH2AX expression was analysed in melanoma cells exposed to H_2O_2 using immunofluorescence. **b**, Microarray analysis of the gene expression profiles of melanoma cells exposed to young/middle and aged fibroblasts identified 63 genes commonly increased in three melanoma cell lines cultured with aged versus young fibroblasts. **c**, Thirty-three genes involved in DNA damage response were significantly altered because of effects of ageing microenvironment in three different melanoma cell lines. Colour scale indicates expression levels relative to aged group. Y, young; M, middle; A, aged.



Extended Data Figure 7 | DNA damage response is increased in melanoma cells exposed to aged fibroblasts. Skin reconstructs made with young or aged fibroblasts were stained with (**a**) anti-53BP1 or (**b**) γH2AX and analysed by immunofluorescence. **c**, FS5 melanoma cell line treated with conditioned media from young and aged fibroblasts showing DNA damage as measured by a comet assay. Two-tailed unpaired *t*-test with Welch's correction was performed between young and aged treatments ($P = 0.039$). Data are represented as mean ± s.e.m. **d**, Percentage ROS activity remaining after NAC treatment of aged fibroblasts. Spearman's correlation between dose and percentage inhibition is significant ($P = 0.043$, $r = -0.700$). Data are represented as mean ± s.d. **e**, Knockdown of SOD3 in young fibroblasts as analysed by western blotting. **f**, Young fibroblasts (2003,071-032 and AG11732) were treated with rsFRP2 (200 ng ml^{-1}) for 72 h and this conditioned medium was used to treat melanoma cells for 48 h. Cells were assessed for DNA damage by γH2AX. **g**, Aged fibroblasts (AG13004 and AG11726) were treated with α-sFRP2 (15 μg ml^{-1}) for 72 h and this conditioned medium was used to treat melanoma cells for 48 h. Cells were assessed for DNA damage by γH2AX.



Extended Data Figure 8 | Analysis of sFRP2, β-catenin, MITF, 8-oxo-dG, APE1, and 53BP1 in individual patients. a, Multiple melanoma samples from aged patients (red bars), and young patients (black bars) were compared. Bars represent average staining intensity (3, highest; 0, lowest) in all patients ($n = 9$ per group) for indicated proteins. **b**–**g**, Dot plots of staining intensity (0–3+) in individual patient samples for (**b**) sFRP2, (**c**) β-catenin, (**d**) MITF, (**e**) APE1, (**f**) 8-oxodG, and (**g**) 53BP1. *P* values for each graph obtained by Mann–Whitney tests.



Extended Data Figure 9 | β-Catenin predicts for sensitivity to vemurafenib. a, Melanoma spheroids were embedded in collagen and treated with 1 μM PLX4720 in the presence of conditioned media from either young or aged fibroblasts. After 48 h, spheroids were assessed for cell death by staining with ethidium homodimer (original magnification ×40). **b**, In cells intrinsically sensitive to vemurafenib in culture, β-catenin expression is increased. **c**, Knockdown of β-catenin in Yumm1.7 cells decreases their sensitivity to PLX4720. Spearman's correlation between dose and percentage proliferation is significant in control cells ($P < 0.0001$, $r = -1.000$) whereas shCTNNB1 cells indicated no significant changes in curve after treatment ($P = 0.948$, $r = 0.03$). **d**, Young mice (8 weeks, $n = 10$ per group) were injected with rsFRP2 (200 ng ml^{-1}, twice weekly) and skin was examined for β-catenin levels by immunohistochemistry. Representative images, original magnification ×200. **e**, Yumm1.7 tumours were injected in aged mice pre-treated with α-sFRP2 antibody (1 mg kg^{-1}, once weekly). Mice were then administered either control or 417 mg kg^{-1} PLX4720-laced chow. ANOVA is significant between treatments ($P < 0.0001$). Two-tailed unpaired *t*-test using rank sum (Mann–Whitney) was performed on tumour volumes on day 25 (1 week after treatment). Results were significant in sFRP2 treatment ($P = 0.036$) and insignificant in IgG2aκ treatment ($P = 0.057$). **f**, Patient samples show a continuum of decreased response in relation to age, Spearman's correlation between percentage response and age is significant ($r = 0.243$, $P = 0.035$). Data are represented as mean ± s.d. for each graph (**c**, **e**).

Extended Data Table 1 | Details of skin fibroblasts obtained from healthy donors

Line	Sex	Age	Disease State	Tissue Type	Biopsy Source
GM04390	F	23	Healthy	Skin	Arm
GM02674	F	29	Healthy	Skin	unspecified
AG11732	F	24	Healthy	Skin	Arm
AG07720	F	24	Healthy	Skin	Arm
AG11735	F	26	Healthy	Skin	Arm
AG07124	F	26	Healthy	Skin	Arm
TP-113	M	27	Healthy	Skin	Arm
2003-071-056	M	32	Healthy	Skin	Arm
GM02673	M	33	Healthy	Skin	unspecified
2003-071-051	M	23	Healthy	Skin	Arm
AG04062	M	31	Healthy	Skin	Arm
2003-071-032	M	28	Healthy	Skin	Arm
2003-071-057	M	31	Healthy	Skin	Arm
AG08620	F	64	Healthy	Skin	Arm
AG11726	F	56	Healthy	Skin	Arm
AG08517	F	66	Healthy	Skin	Arm
AG11489	F	66	Healthy	Skin	Arm
AG07757	F	61	Healthy	Skin	Arm
AG08529	F	60	Healthy	Skin	Arm
AG08701	F	62	Healthy	Skin	Arm
AG08379	F	60	Healthy	Skin	Arm
AG12597	F	62	Healthy	Skin	Arm
AG12988	F	56	Healthy	Skin	Arm
AG12589	F	68	Healthy	Skin	Arm
AG09878	F	61	Healthy	Skin	Arm
GM13335	M	57	Healthy	Skin	Arm
AG13095	M	62	Healthy	Skin	Arm
AG09157	M	63	Healthy	Skin	Arm
AG11079	M	63	Healthy	Skin	Arm
AG10942	M	50	Healthy	Skin	Arm
AG06281	M	67	Healthy	Skin	Arm
AG13004	M	68	Healthy	Skin	Arm

LETTER

doi:10.1038/nature17393

Reductive carboxylation supports redox homeostasis during anchorage-independent growth

Lei Jiang[1], Alexander A. Shestov[2], Pamela Swain[3], Chendong Yang[1], Seth J. Parker[4], Qiong A. Wang[5], Lance S. Terada[6], Nicholas D. Adams[7], Michael T. McCabe[7], Beth Pietrak[7], Stan Schmidt[7], Christian M. Metallo[4], Brian P. Dranka[3], Benjamin Schwartz[7] & Ralph J. DeBerardinis[1,8,9]

Cells receive growth and survival stimuli through their attachment to an extracellular matrix (ECM)[1]. Overcoming the addiction to ECM-induced signals is required for anchorage-independent growth, a property of most malignant cells[2]. Detachment from ECM is associated with enhanced production of reactive oxygen species (ROS) owing to altered glucose metabolism[2]. Here we identify an unconventional pathway that supports redox homeostasis and growth during adaptation to anchorage independence. We observed that detachment from monolayer culture and growth as anchorage-independent tumour spheroids was accompanied by changes in both glucose and glutamine metabolism. Specifically, oxidation of both nutrients was suppressed in spheroids, whereas reductive formation of citrate from glutamine was enhanced. Reductive glutamine metabolism was highly dependent on cytosolic isocitrate dehydrogenase-1 (IDH1), because the activity was suppressed in cells homozygous null for IDH1 or treated with an IDH1 inhibitor. This activity occurred in absence of hypoxia, a well-known inducer of reductive metabolism. Rather, IDH1 mitigated mitochondrial ROS in spheroids, and suppressing IDH1 reduced spheroid growth through a mechanism requiring mitochondrial ROS. Isotope tracing revealed that in spheroids, isocitrate/citrate produced reductively in the cytosol could enter the mitochondria and participate in oxidative metabolism, including oxidation by IDH2. This generates NADPH in the mitochondria, enabling cells to mitigate mitochondrial ROS and maximize growth. Neither IDH1 nor IDH2 was necessary for monolayer growth, but deleting either one enhanced mitochondrial ROS and reduced spheroid size, as did deletion of the mitochondrial citrate transporter protein. Together, the data indicate that adaptation to anchorage independence requires a fundamental change in citrate metabolism, initiated by IDH1-dependent reductive carboxylation and culminating in suppression of mitochondrial ROS.

In monolayer cultures, growth factors direct cells to take up glucose and glutamine and use them to produce macromolecules. Both nutrients are used to produce the lipogenic precursor citrate (Extended Data Fig. 1a). To identify metabolic alterations during anchorage independence, H460 lung cancer cells were detached from monolayers and aggregated into spheroids. Cells within spheroids proliferated at a reduced rate (Extended Data Fig. 2a). Although growth in both conditions required glucose and glutamine (Extended Data Fig. 2b), spheroids consumed less of both and secreted less lactate, glutamate and ammonia (Extended Data Fig. 2c, d). The ratio of ammonia released to glutamine consumed was comparable between conditions (Extended Data Fig. 2d). Spheroids displayed reduced entry of glucose-derived carbon into citrate (Fig. 1a) and consumed less oxygen per cell (Fig. 1b). These findings implied reduced pyruvate dehydrogenase (PDH) activity, as demonstrated previously during matrix detachment[3]. Indeed, inhibitory PDH phosphorylation and expression of PDH kinase-1 (PDK1) were elevated in spheroids (Fig. 1c). Citrate labelling from [U-^{13}C]glutamine persisted in spheroids, but the ^{13}C distribution was altered, particularly in that the m+5 fraction (the fraction containing five ^{13}C nuclei) exceeded m+4 (Fig. 1d). This persisted when cells were disaggregated and permitted to reform spheroids (Extended Data Fig. 2e). The m+5 fraction appeared rapidly and endured as the most prominent labelled form (Fig. 1e), regardless of the type of culture medium (Supplementary Table 1; this Table contains all ^{13}C data throughout the paper). Because PDH inhibition can alter glutamine

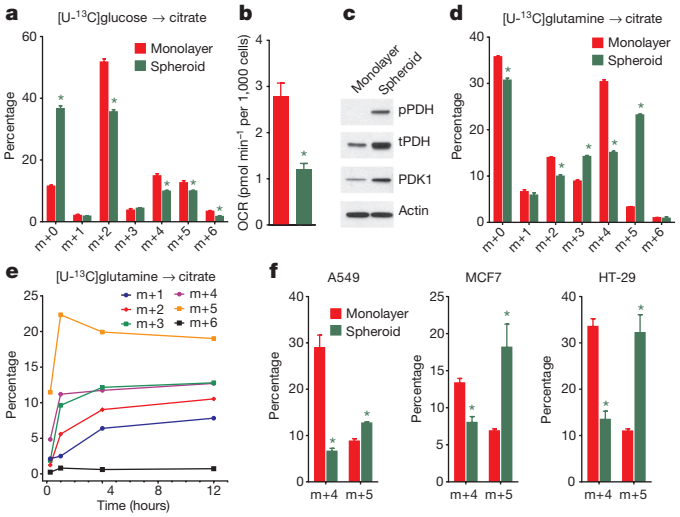


Figure 1 | Reductive glutamine metabolism in spheroids. a, Mass isotopologue analysis of citrate in H460 cells cultured with [U-^{13}C]glucose and unlabelled glutamine ($n = 3$ cultures from a representative experiment). **b**, Oxygen consumption rates (OCR) of cells grown in monolayer or spheroid culture ($n = 10$ monolayer cultures and 11 spheroids from a representative experiment). **c**, Western blot for total (t) and phosphorylated (p, Ser 293) PDH, and PDH kinase-1 (PDK1). **d**, Mass isotopologue analysis of citrate in cells cultured with [U-^{13}C]glutamine and unlabelled glucose ($n = 3$ cultures from a representative experiment). **e**, Evolution of citrate mass isotopologues in spheroids cultured with [U-^{13}C]glutamine ($n = 2$ cultures for each time point). **f**, Citrate m+4 and m+5 isotopologues in monolayer and spheroid cultures of A549, HT-29 and MCF7 cells cultured with [U-^{13}C]glutamine ($n = 3$ A549 monolayer cultures; $n = 4$ cultures for all other conditions). Complete mass isotopologue distributions are shown in Supplementary Table 1. All data represent mean ± s.d. $^*P < 0.05$, Welch's unequal variances t-test. All experiments were repeated 3 times or more.

[1]Children's Medical Center Research Institute, UT Southwestern Medical Center, Dallas, Texas 75390-8502, USA. [2]Department of Radiology, University of Pennsylvania School of Medicine, 3620 Hamilton Walk, Philadelphia, Pennsylvania 19104, USA. [3]Seahorse Bioscience, 16 Esquire Road, North Billerica, Massachusetts 01862, USA. [4]Department of Bioengineering, University of California, San Diego, La Jolla, California 92093, USA. [5]Touchstone Diabetes Center, UT Southwestern Medical Center, Dallas, Texas 75390, USA. [6]Department of Internal Medicine, UT Southwestern Medical Center, Dallas, Texas 75390, USA. [7]GlaxoSmithKline, 1250 South Collegeville Road, Collegeville, Pennsylvania 19426, USA. [8]Department of Pediatrics, UT Southwestern Medical Center, Dallas, Texas 75390, USA. [9]McDermott Center for Human Growth and Development, UT Southwestern Medical Center, Dallas, Texas 75390, USA.

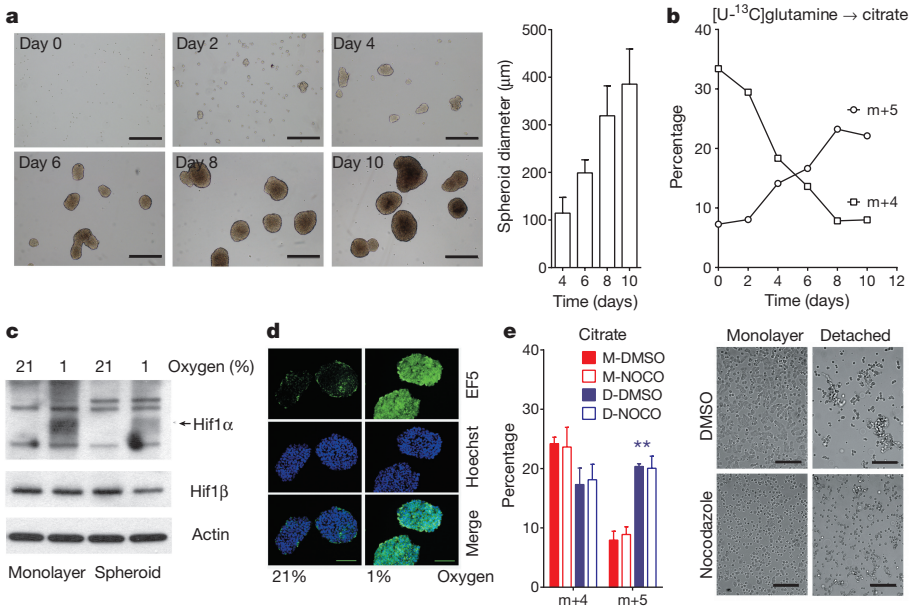


Figure 2 | Spheroid metabolism is distinct from the metabolic alterations induced by hypoxia. a, Time-dependent increase in H460 spheroid diameter. Scale bars, 500 μm. ($n = 20$ spheroids on days 4 and 6; $n = 18$ spheroids on day 8; $n = 16$ spheroids on day 10). **b**, Citrate isotopologues in spheroids cultured for 4 h with [U-^{13}C]glutamine and unlabelled glucose ($n = 2$ cultures for each time point). **c**, Effect of hypoxia on HIF1α protein expression in monolayer and spheroid culture. HIF1α protein is indicated by an arrow, with nonspecific bands also present on the blot. **d**, EF5 staining in sectioned spheroids cultured under 21% or 1% oxygen for 16 h. Scale bars, 100 μm. **e**, Citrate m+4 and m+5 isotopologues in monolayer (M) or detached (D) H460 cells labelled with [U-^{13}C]glutamine and treated with DMSO or nocodazole (NOCO) to prevent aggregation ($n = 4$ cultures from two experiments). Scale bars, 200 μm. All data represent mean ± s.d. *$P < 0.05$, Welch's unequal variances t-test followed by multiple-comparison correction. Experiments in **a** and **b** were repeated twice. Experiments in **c**, **d** and **e** were repeated 3 times or more.

metabolism[4], we examined the effect of the PDK1 inhibitor dichloroacetate (DCA), which activates PDH, on ^{13}C labelling. DCA enhanced glucose-dependent citrate labelling and reduced the m+5 fraction from [U-^{13}C]glutamine (Extended Data Fig. 2f), indicating that m+5 citrate resulted from reduced PDH activity.

Culture with [1-^{13}C]glutamine demonstrated that spheroids induced reductive glutamine metabolism to generate isocitrate/citrate (Extended Data Fig. 3a). Reductive citrate labelling was observed in spheroids from multiple lung, colon and breast cancer cell lines (Fig. 1f). However, labelling of other TCA cycle intermediates predominantly reflected oxidative (m+4) rather than reductive (m+3) metabolism (Extended Data Fig. 3b). To test whether reductive metabolism occurred in non-transformed cells, we compared [U-^{13}C]glutamine metabolism between lung cancer cells and nonmalignant bronchial epithelial cells (BECs) from the same patient[5]. Cancer cells but not BECs displayed enhanced citrate m+5 labelling upon detachment (Extended Data Fig. 3c).

Reductive carboxylation is enhanced during hypoxia through a HIF1-dependent mechanism that transmits glutamine carbon to fatty acids[6]. Although large spheroids contain gradients of oxygenation, reductive labelling occurred in spheroids much smaller than the limit of oxygen diffusion[7] (Fig. 2a, b), and hyperoxia did not normalize citrate m+5 (Extended Data Fig. 4a). We detected neither HIF1α

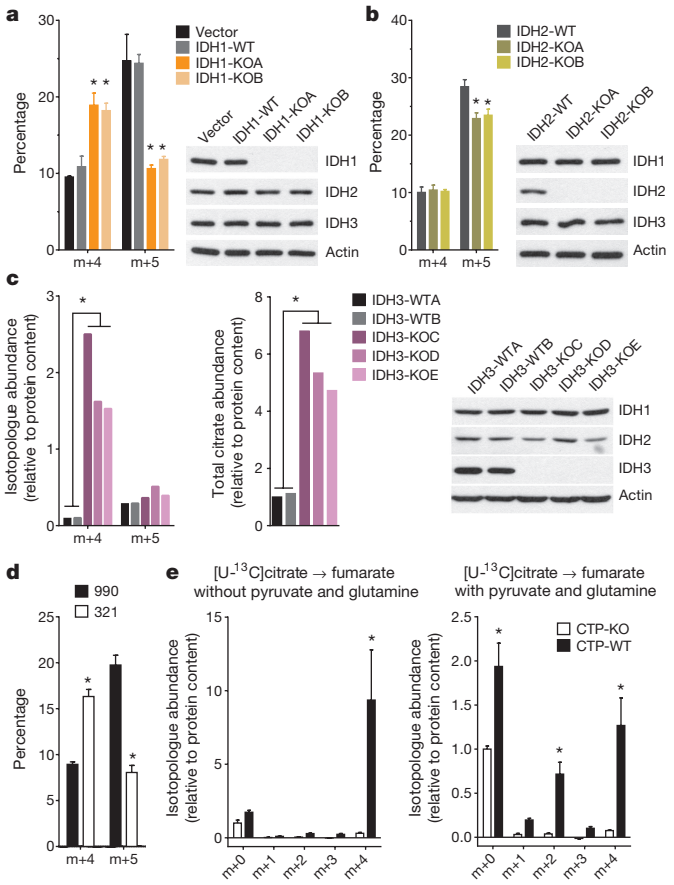


Figure 3 | Reductive carboxylation in spheroids is primarily dependent on cytosolic IDH1. a, **b**, Fractional abundance of citrate m+4 and m+5 in spheroids containing or lacking IDH1 (**a**) or IDH2 (**b**) cultured in [U-^{13}C]glutamine ($n = 3$ cultures from two experiments). **c**, Abundance of total citrate and of citrate m+4 and m+5 from [U-^{13}C]glutamine in H460 spheroids containing or lacking IDH3 ($n = 2$ cultures from a representative experiment). **d**, Fractional abundance of citrate m+4 and m+5 in H460 spheroids cultured with [U-^{13}C]glutamine and treated with an IDH1 inhibitor (321) or a structurally similar control compound (990) ($n = 4$ cultures from two experiments). **e**, Fumarate mass isotopologues in mitochondria isolated from cells containing or lacking the citrate transporter protein (CTP-WT and CTP-KO, respectively). Isolated mitochondria were cultured with [U-^{13}C] citrate as the sole carbon source (left), or [U-^{13}C]citrate with unlabelled pyruvate and glutamine (right) ($n = 4$ cultures from two experiments). All data represent mean ± s.d. *$P < 0.05$, ANOVA (**a**, **b**), or Welch's unequal variances t-test (**c**–**e**). All experiments were repeated 3 times or more.

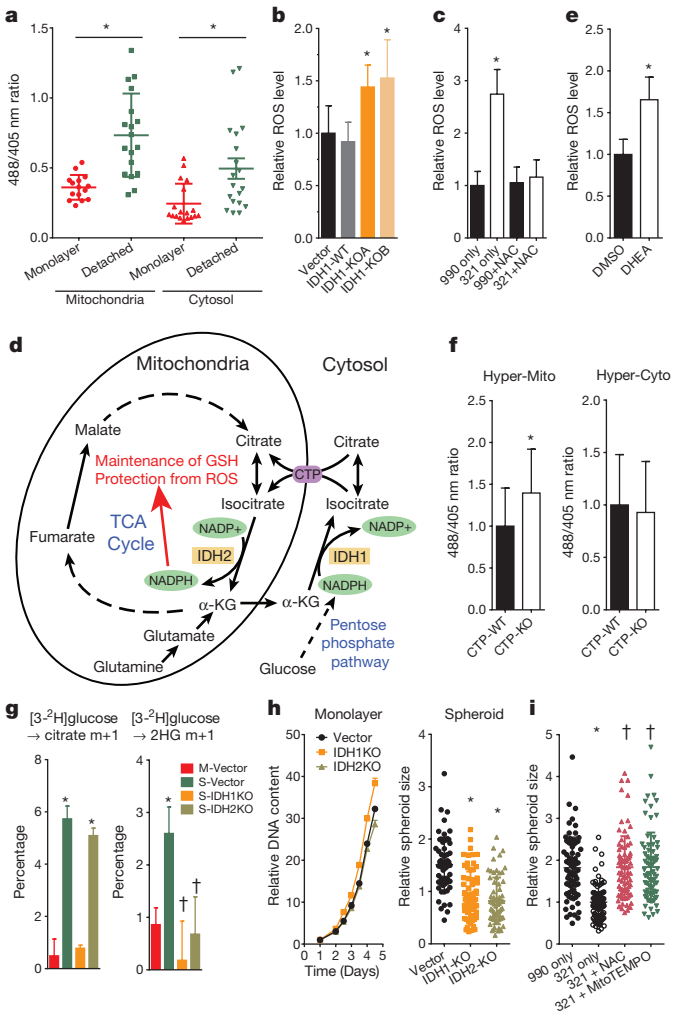


Figure 4 | Reductive glutamine metabolism mitigates mitochondrial ROS and promotes spheroid growth. a, Transient expression of mitochondrial and cytosolic hydrogen peroxide reporter (HyPer) vectors in H460 cells. Cells were cultured under monolayer or detached conditions for 48 h before imaging to assess ROS levels. ($n = 15$ HyperMito monolayer cells; $n = 18$ HyperMito detached cells; $n = 18$ HyperCyto monolayer cells; $n = 19$ HyperCyto detached cells). **b**, Staining of spheroids containing or lacking IDH1 with the mitochondrial ROS sensor MitoSOX. ($n = 16$ Vector spheroids; $n = 18$ IDH1-WT spheroids; $n = 17$ IDH1-KOA spheroids; $n = 15$ IDH1-KOB spheroids). **c**, Effect of IDH1 inhibition by compound 321 and the ROS scavenger *N*-acetylcysteine (NAC) on MitoSOX signal in H460 spheroids. ($n = 31$ compound 990 only spheroids; $n = 25$ compound 321 only spheroids; $n = 33$ compound 990+NAC spheroids; $n = 30$ compound 321+NAC spheroids). **d**, Pathway through which cytosolic reductive carboxylation, followed by isocitrate/citrate import into the mitochondria, can contribute to the mitochondrial NADPH pool and suppress mitochondrial ROS. α-KG, α- ketoglutarate. **e**, MitoSOX staining of DHEA-treated H460 spheroids ($n = 53$ spheroids for each condition). **f**, HyPer signal for mitochondrial (left) and cytosolic (right) ROS in H460 cells containing or lacking CTP. ($n = 28$ HyperCyto CTP-KO cells; $n = 35$ cells for all other conditions). **g**, Labelling from [3-^{2}H]glucose in citrate (left) or 2-HG (right) in H460-derived cells expressing mutant IDH2 and containing or lacking wild type IDH1 and IDH2 ($n = 3$ cultures, $^*P < 0.05$, comparing to M-Vector, $\dagger P < 0.05$, comparing to S-Vector). M, monolayer; S, spheroid. **h**, Monolayer and spheroid growth of H460 cells containing or lacking IDH1 or IDH2. ($n = 6$ monolayer cultures for each cell line; $n = 59$ Vector spheroids; $n = 70$ IDH1-KO spheroids; $n = 65$ IDH2-KO spheroids). **i**, Spheroid size after 8 days of compound 321 treatment, with or without NAC or the mitochondrial ROS scavenger MitoTEMPO. ($n = 75$ compound 990 only spheroids; $n = 77$ compound 321 only spheroids; $n = 81$ compound 321+NAC spheroids; $n = 77$ compound 321+MitoTEMPO spheroids, $\dagger P < 0.05$ comparing to treatment with compound 321 only). $^*,\dagger P < 0.05$, ANOVA (**b**, **c**, **g**, **h** and **i**), or Welch's unequal variances *t*-test (**a**, **e** and **f**). All data represent mean ± s.d. Experiments in **f** were repeated twice, and all other experiments were repeated 3 times or more.

stabilization nor staining with a hypoxia probe in spheroids cultured under 21% oxygen (Fig. 2c, d). Furthermore, although large spheroids contain gradients of nutrient availability[8], experimentally reducing glucose/glutamine availability did not increase citrate m+5 (Extended Data Fig. 4b). Most compellingly, detachment without aggregation was sufficient to enhance citrate m+5 (Fig. 2e), and spheroids lost the reductive pattern when allowed to adhere to plastic (Extended Data Fig. 4c). Hypoxia elicited numerous labelling changes distinct from patterns observed in normoxic spheroids (Extended Data Fig. 5a and Supplementary Discussion). Remarkably, reductive citrate labelling was not associated with increased contribution of glutamine to palmitate unless the spheroids were cultured under hypoxia (Extended Data Fig. 5b). Glucose was the predominant lipogenic carbon source in both monolayer and spheroid cells (Extended Data Fig. 5c). Acetate, a source of lipogenic acetyl-CoA in hypoxia[9], was only a minor source of lipogenic carbon in normoxic H460 spheroids (Extended Data Fig. 5d). Together, these data suggest that anchorage loss per se rather than oxygen/nutrient limitation stimulates a mode of reductive metabolism distinct from hypoxia.

We next used metabolic flux analysis to better understand the metabolic effects of anchorage loss. Steady state fluxes were calculated by integrating extracellular flux rates and ^{13}C distributions in several metabolites from multiple tracers (see Methods and Supplementary Tables 2 and 3). Modelling with a set of conventional reactions and compartmentation did not produce an adequate fit for the spheroid data (Extended Data Fig. 6a). This was related to discrepant labelling between citrate, which contained evidence of enhanced reductive carboxylation, and palmitate, which did not. To rectify this discrepancy, we modified the model to force isocitrate/citrate produced reductively in the cytosol into the mitochondria, as if by channelling (Extended Data Fig. 6b). This pool then mixed with mitochondrial isocitrate/citrate before being used for fatty acid synthesis in the cytosol. This modification greatly improved the model's ability to fit the data, with the resulting fit indicating persistent but reduced glycolysis, PDH flux and glucose/glutamine oxidation in spheroids, as well as an enhanced reductive IDH flux in the cytosol (Extended Data Fig. 6b and Extended Data Table 1). The model also indicated bidirectional isocitrate/citrate traffic across the mitochondrial membrane, with net flux in the direction of export to the cytosol.

Mammals contain three IDH isoforms. IDH1 and IDH2 catalyse the reversible, NADP$^+$/NADPH-dependent interconversion of isocitrate and α-ketoglutarate in the cytosol and mitochondria, respectively, and participate in reductive carboxylation[10]. IDH3 is mitochondrial, NAD$^+$-dependent and essentially irreversible[11]. To determine the roles of all three isoforms in spheroids, CRISPR/Cas9 was used to generate H460 clones lacking each one. IDH1 or IDH2 knockout reduced citrate m+5, although the effect of IDH1 was more pronounced (Fig. 3a, b). Deleting IDH3 caused a large accumulation of citrate (Fig. 3c), whereas deleting IDH1 or IDH2 had only minor effects on citrate abundance (Extended Data Fig. 7a). Thus, IDH3 is primarily responsible for citrate oxidation in the TCA cycle, whereas IDH1 is primarily responsible for reductive carboxylation. Like IDH1 deletion, an IDH1 inhibitor (compound 321)[12] eliminated the spheroids' gain in citrate m+5 and enhanced the m+4 fraction (Fig. 3d and Extended Data Fig. 7b–e). The drug had no effect on citrate m+5 in IDH1-deleted spheroids, consistent with its selectivity for IDH1 (Extended Data Fig. 7e).

Because reductive citrate metabolism was not associated with enhanced palmitate labelling from glutamine, we tested whether citrate was transported into mitochondria, as predicted by the model. Citrate export to the cytosol through a citrate transporter protein (CTP) encoded by *SLC25A1* is a key component of lipogenesis[13], but

mitochondrial citrate import has also been reported[14]. We prepared mitochondria from H460 cells containing or lacking CTP and examined their ability to metabolize citrate (Extended Data Fig. 8a, b). CTP-wild type mitochondria took up [U-^{13}C]citrate and converted it to succinate, fumarate and malate, with the majority of these intermediates labelled as m+4 in the first turn of the TCA cycle (Fig. 3e and Extended Data Fig. 8c). Adding unlabelled glutamine and pyruvate to [U-^{13}C]citrate resulted in m+2 labelling from multiple turns (Fig. 3e and Extended Data Fig. 8d). Labelling from [U-^{13}C]citrate was essentially absent in CTP-deficient mitochondria, even though [U-^{13}C]glutamine was metabolized normally (Fig. 3e and Extended Data Fig. 8c–e).

In non-transformed breast epithelial cells, matrix detachment enhances ROS by reducing pentose phosphate pathway (PPP) activity, but oncogenes sustain the PPP and viability during detachment[2]. Transgenic reporters that sense ROS in either the cytosol or mitochondria[15] revealed that detachment increased ROS in both compartments, but particularly the mitochondria (Fig. 4a). IDH1 knockout (Fig. 4b) or inhibition with 321 (Fig. 4c) further increased mitochondrial ROS, consistent with a pathway in which isocitrate/citrate is formed through NADPH-dependent reductive carboxylation in the cytosol, followed by metabolism of one or both metabolites in the mitochondria to produce NADPH and fortify ROS defences (Fig. 4d and Extended Data Fig. 1b). Consistent with this model, both the PPP inhibitor dehydroepiandrosterone (DHEA) and IDH2 knockout enhanced mitochondrial ROS (Fig. 4e and Extended Data Fig. 9a). IDH1 knockout reduced cytosolic ROS, suggesting that PPP-derived NADPH is repurposed to counteract cytosolic ROS in the absence of IDH1 (Extended Data Fig. 9b). CTP knockout enhanced mitochondrial ROS without affecting cytosolic ROS, as predicted if mitochondrial isocitrate/citrate import generates reducing equivalents to mitigate mitochondrial ROS (Fig. 4f).

To examine reducing equivalent metabolism, H460 cells were cultured with [3-^{2}H]glucose, which labels cytosolic NADPH via the PPP, but does not transfer ^{2}H to TCA cycle metabolites in monolayer culture[16]. We generated parental, IDH1- and IDH2-knockout H460 cells expressing a mutant IDH2 that produces 2-hydroxyglutarate (2HG) in an NADPH-dependent mitochondrial reaction. Thus, ^{2}H labelling of 2HG reflects mitochondrial NADPH labelling[16]. As expected, citrate was not labelled from [3-^{2}H]glucose in monolayer cells. However, spheroids contained substantial citrate labelling that required IDH1 but not IDH2 (Fig. 4g) and was suppressed by DHEA (Extended Data Fig. 9c). Similarly, 2HG was essentially unlabelled in monolayer cells, but labelling was evident in spheroids; here, labelling required both IDH1 and IDH2 (Fig. 4g). All these data are consistent with the pathway in Fig. 4d.

Finally, CTP-, IDH1- and IDH2-deficient cells were challenged to grow as spheroids. CTP deletion suppressed growth in both monolayer and spheroids, reflecting the transporter's dual roles in lipogenesis and ROS mitigation (Extended Data Fig. 9d). By contrast, IDH1 and IDH2 were dispensable in monolayers, but loss of either enzyme suppressed average spheroid size by half (Fig. 4h) and growth could be partially rescued with the mitochondrial ROS scavenger MitoTEMPO (Extended Data Fig. 9e). Treatment with 321 also reduced spheroid size, and this was completely rescued by MitoTEMPO or *N*-acetylcysteine (NAC) (Fig. 4i).

Protection against oxidative stress is thought to be one aspect of metabolic reprogramming that supports cancer cell fitness[17–19]. Anchorage independence induces additional oxidative stress resulting in death unless NADPH-producing pathways are engaged[2]. The finding that cytosolic IDH1 is required to regulate mitochondrial ROS was surprising. IDH1 and IDH2 participate in a cycle transmitting NADPH from mitochondria to cytosol[20]. Our data suggest that this pathway also operates in reverse to transfer reducing equivalents from the PPP into the mitochondria. This pathway enables an alternative form of redox regulation and suggests a possible metabolic intervention to suppress anchorage independence. The importance of this pathway for redox homeostasis is remarkable, considering that the net direction of citrate trafficking in spheroids is from the mitochondria to the cytosol to supply fatty acid synthesis. The isocitrate/citrate fluxes responsible for ROS mitigation may arise from metabolite compartmentation involving mechanisms yet to be characterized. The data imply that limiting mitochondrial rather than cytosolic oxidative stress is crucial for anchorage independence, because IDH1 loss enhanced the former while reducing the latter, culminating in reduced spheroid size.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 23 January 2015; accepted 2 February 2016.
Published online 6 April 2016.

1. Bhowmick, N. A., Neilson, E. G. & Moses, H. L. Stromal fibroblasts in cancer initiation and progression. *Nature* **432,** 332–337 (2004).
2. Schafer, Z. T. *et al.* Antioxidant and oncogene rescue of metabolic defects caused by loss of matrix attachment. *Nature* **461,** 109–113 (2009).
3. Grassian, A. R., Metallo, C. M., Coloff, J. L., Stephanopoulos, G. & Brugge, J. S. Erk regulation of pyruvate dehydrogenase flux through PDK4 modulates cell proliferation. *Genes Dev.* **25,** 1716–1733 (2011).
4. Rajagopalan, K. N. *et al.* Metabolic plasticity maintains proliferation in pyruvate dehydrogenase deficient cells. *Cancer Metab.* **3,** 7 (2015).
5. Kim, H. S. *et al.* Systematic identification of molecular subtype-selective vulnerabilities in non-small-cell lung cancer. *Cell* **155,** 552–566 (2013).
6. Metallo, C. M. *et al.* Reductive glutamine metabolism by IDH1 mediates lipogenesis under hypoxia. *Nature* **481,** 380–384 (2012).
7. Friedrich, J., Seidel, C., Ebner, R. & Kunz-Schughart, L. A. Spheroid-based drug screen: considerations and practical approach. *Nature Protocols* **4,** 309–324 (2009).
8. Hunnewell, M. G. & Forbes, N. S. Active and inactive metabolic pathways in tumor spheroids: determination by GC-MS. *Biotechnol. Prog.* **26,** 789–796 (2010).
9. Schug, Z. T. *et al.* Acetyl-CoA synthetase 2 promotes acetate utilization and maintains cancer cell growth under metabolic stress. *Cancer Cell* **27,** 57–71 (2015).
10. Mullen, A. R. *et al.* Reductive carboxylation supports growth in tumour cells with defective mitochondria. *Nature* **481,** 385–388 (2012).
11. Garrett, R. H. & Grisham, C. M. *Biochemistry* 618 (Brooks Cole, 2004).
12. Okoye-Okafor, U. C. *et al.* New IDH1 mutant inhibitors for treatment of acute myeloid leukemia. *Nature Chem. Biol.* **11,** 878–886 (2015).
13. Gnoni, G. V., Priore, P., Geelen, M. J. & Siculella, L. The mitochondrial citrate carrier: metabolic role and regulation of its activity and expression. *IUBMB Life* **61,** 987–994 (2009).
14. Kaplan, R. S., Morris, H. P. & Coleman, P. S. Kinetic characteristics of citrate influx and efflux with mitochondria from Morris hepatomas 3924A and 16. *Cancer Res.* **42,** 4399–4407 (1982).
15. Belousov, V. V. *et al.* Genetically encoded fluorescent indicator for intracellular hydrogen peroxide. *Nature Methods* **3,** 281–286 (2006).
16. Lewis, C. A. *et al.* Tracing compartmentalized NADPH metabolism in the cytosol and mitochondria of mammalian cells. *Mol. Cell* **55,** 253–263 (2014).
17. Anastasiou, D. *et al.* Inhibition of pyruvate kinase M2 by reactive oxygen species contributes to cellular antioxidant responses. *Science* **334,** 1278–1283 (2011).
18. Piskounova, E. *et al.* Oxidative stress inhibits distant metastasis by human melanoma cells. *Nature* **527,** 186–191 (2015).
19. Jeon, S.-M., Chandel, N. S. & Hay, N. AMPK regulates NADPH homeostasis to promote tumour cell survival during energy stress. *Nature* **485,** 661–665 (2012).
20. Sazanov, L. A. & Jackson, J. B. Proton-translocating transhydrogenase and NAD- and NADP-linked isocitrate dehydrogenases operate in a substrate cycle which contributes to fine regulation of the tricarboxylic acid cycle activity in mitochondria. *FEBS Lett.* **344,** 109–116 (1994).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank M. Mitsche for help with analysis of lipogenic acetyl-CoA enrichment, A. Grassian for advice about IDH1, J. Garcia for hypoxia/hyperoxia experiments and C. Frezza for discussion of mitochondrial isolation. J. Kozlitina assisted with statistical analysis and R. Egnatchik provided advice about metabolic flux analysis. R.J.D. is supported by grants from the N.I.H (R01CA157996), Cancer Prevention and Research Institute of Texas (RP130272) and Robert A. Welch Foundation (I1733). C.M.M. is supported by N.I.H. grant R01CA188652.

Author Contributions L.J. and R.J.D. designed the study. L.J., Q.A.W. and C.Y. performed molecular and cell biology experiments. P.S. and B.P.D. performed Seahorse experiments. L.S.T., S.J.P. and C.M.M. provided reagents and expertise for ROS and ^{2}H tracing experiments. L.J. and A.A.S. performed metabolic flux analysis. N.D.A., M.T.M., B.P., S.S. and B.S. provided the IDH1 inhibitor. L.J. and R.J.D. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.J.D. (Ralph.Deberardinis@UTSouthwestern.edu).

METHODS

Reagents. MitoSOX Red mitochondrial superoxide indicator was from Life Technologies. MitoTEMPO, *N*-acetylcysteine (NAC), Nocodazole, DHEA and DCA were from Sigma. The IDH1 inhibitor 321 and control compound 990 were from GlaxoSmithKline[12].

Cell lines and culture. Cell lines were identified using DNA fingerprinting and confirmed to be mycoplasma free. H460, A549, HBEC30 and HCC4017 cells were provided by J. D. Minna, UT Southwestern. MCF7 cells were provided by David A. Boothman, UT Southwestern. HT-29 cells were provided by Jared Rutter, University of Utah. H460 and A549 cells were cultured in RPMI supplemented with penicillin/streptomycin, 5% fetal bovine serum (FBS), L-glutamine (4 mM) and 1 mM HEPES. HT-29 and MCF7 were cultured in DMEM supplemented with penicillin/streptomycin, 10% fetal bovine serum (FBS), L-glutamine (4 mM) and 1 mM HEPES. HBEC30 and HCC4017 cells were cultured in defined medium[5]. Dishes with an Ultra-Low Attachment surface were used for suspension and spheroid culture. Identical culture medium was used for monolayer and spheroid culture. For spheroids, 2×10^5 H460 cells were plated in a 10 cm Ultra-Low Attachment dish. The medium was changed on days 4, 6 and 7 of culture, by centrifuging at 50*g* for 3 min, then gently resuspending in fresh medium. For labelling assays, the spheroids were resuspended in medium containing ^{13}C-labelled nutrients, and the time of resuspension was considered time 0 of the labelling period.

CRISPR/Cas9-mediated recombination. *IDH1*, *IDH2*, *IDH3* and *SLC25A1*-deficient H460 cell lines were generated using the CRISPR/Cas9 system[21]. Wild type clones were selected from both the control vector (Vector) and targeting vector transfections (WT). In order to control for variations among individual clones, 4 to 5 clones were pooled together, and different pools for each targeted gene were used for further experiments.

Cell growth assays. Cell proliferation was measured by counting cells after trypsinization. DNA content was also used to monitor monolayer cell growth, as previously described[22]. Spheroid size was determined by measuring the maximum cross-sectional area of individual spheroids using ImageJ software.

Metabolic assays and stable isotope tracing in intact cells. Glucose, lactate, glutamine and glutamate were measured in culture medium using an automated electrochemical analyser (BioProfile Basic-4 analyser, NOVA). Ammonia was measured using an enzymatic assay (Megazyme). Nutrients labelled with ^{13}C, ^{2}H or ^{15}N were purchased from Cambridge Isotope Laboratories. Stable isotope tracing experiments to determine isotopologue distributions in soluble metabolites and fatty acids were performed as described previously[22,23]. For deuterium tracing, H460 clones were engineered to express the IDH2-R172K mutant (mtIDH2) under control of doxycycline[16]. Clones lacking wild-type IDH1 or IDH2, or containing both, were generated. Spheroids were cultivated for 7 days, then doxycycline (0.2 μg ml^{-1}) was added for 24 h to induce mtIDH2. On day 8, spheroids were cultured in RPMI containing 10 mM [3-^{2}H]glucose and unlabelled glutamine for 6 h.

Mitochondria isolation and isotope tracing. Mitochondria were prepared with the Qproteome Mitochondria Isolation Kit (Qiagen). Isotope tracing was modified from a previously described procedure[22]. Mitochondrial pellets were reconstituted in assay buffer (125 mM KCl, 10 mM Tris/MOPS, 0.1 mM EGTA/Tris, 1 mM P_i, pH 7.4) supplied with indicated nutrients and tracer. For glutamine tracing, 40 μM [U-^{13}C]glutamine and 40 μM unlabelled pyruvate were added to the assay buffer. For citrate tracing, 40 μM [U-^{13}C]citrate with or without 40 μM unlabelled glutamine and 40 μM unlabelled pyruvate were added to the assay buffer. Mitochondria were incubated in the tracing buffer for 10 min, at 30 °C with 500 r.p.m. agitation in a heat block.

Metabolic flux analysis (MFA). Steady state metabolic fluxes were calculated by combining extracellular flux rates (glucose/glutamine utilization, lactate/alanine/glutamate secretion) and ^{13}C mass isotopologue distributions (MIDs) for citrate, glutamate, fumarate, malate, aspartate and palmitate, using the INCA software package[24], which applies an elementary metabolite unit framework to efficiently simulate MIDs[25,26]. We developed reaction networks describing the stoichiometry and carbon transitions of central carbon metabolism (Supplementary Table 2), with assumptions as previously described[4] and summarized below. Parallel labelling data from cultures fed [1-^{13}C]glutamine, [5-^{13}C]glutamine, [U-^{13}C]glutamine or [U-^{13}C]glucose were used to simultaneously fit the same network model to estimate intracellular fluxes. Data used in metabolic flux analysis for monolayer and spheroid cultures are reported in Supplementary Table 3. To ensure that a global minimum of fluxes was identified, flux estimations were initiated from random values and repeated a minimum of 50 times. A chi-square test was applied to test goodness-of-fit, and accurate 95% confidence intervals were calculated by assessing the sensitivity of the sum of squared residuals to flux parameter variations. Extended Data Table 1 contains the degrees of freedom and sum-of-squared residuals (SSR) for the best fit model and the lower and upper bounds of 95% confidence intervals for all fluxes.

MFA procedures and assumptions:

(1) During the experiments, cells are at metabolic steady state.

(2) $^{13}CO_2$ produced during oxidation reactions is not reincorporated via carboxylation reactions.

(3) Cells are given 24 h to metabolize ^{13}C substrates. After 24 h, it is assumed that the isotopic labelling has reached steady state.

(4) The metabolites succinate and fumarate are symmetrical and their metabolism through the TCA cycle does not produce a particular orientation.

(5) The metabolites pyruvate, acetyl-CoA, citrate, α-ketoglutarate, malate, fumarate and oxaloacetate are metabolically active in both the cytosol and mitochondria. Malate and α-ketoglutarate are allowed to freely mix between the compartments.

(6) During the extraction process, intracellular pools of metabolites are homogenized. Therefore GC-MS analysis of the isotopic enrichment of these metabolites reflects the mixture of distinct metabolic pools. By employing the INCA platform to perform metabolic flux analysis, it is possible to extract meaningful information from these mixed pools. To do this, the model employs parameters to account for the mixing of mitochondrial and cytosolic metabolites.

(7) For flux calculations in spheroids, it is not possible to reconcile the palmitate and citrate labelling from the glutamine tracers by mixing the citrate pool in cytosol. To achieve adequate fits, isocitrate/citrate generated from cytosolic reductive carboxylation is first allowed to mix with mitochondrial citrate. This mixed pool could then be used for oxidation in the TCA cycle and fatty acid synthesis.

Western blotting. Whole cells/spheroids or mitochondrial lysates were prepared in RIPA buffer and quantified using the BCA Protein Assay (Thermo Scientific). Proteins were separated on 4–20% SDS–PAGE gels, transferred to PVDF membranes, and probed with antibodies against IDH1 (ab94571), IDH2 (ab55271), IDH3 (ab58641) from Abcam, PDHα (#459400, Thermo), PDHα-pSer293 (AP1062), GAPDH (AB2302) from Millipore, PDK1 (#3820), Hif1α (#3716) from Cell Signaling, CTP (sc-86392), AIF (sc-13116) from Santa Cruz Biotechnology and Actin (A3853, Sigma).

EF5 staining. Day 7 H460 spheroids were cultured under normoxia (21% oxygen) or hypoxia (1% oxygen) for 16 h. The spheroids were then treated with 100 μM EF5 compound for 3 h, fixed in 4% paraformaldehyde, and embedded in OCT for frozen sectioning. 10 μm spheroid sections were stained with EF5 antibody as previously described[27].

Subcellular ROS detection. Cytosolic and mitochondrial ROS levels were measured with the organelle-specific HyPer system as previously described[15,28]. Briefly, trypsinized H460 cells were transfected with HyPer-cyto or HyPer-mito vectors. Half of the transfected cells were allowed to attach as a monolayer, and the other half were cultured in suspension using Ultra-Low Attachment dishes. Images were acquired 48 h after transfection, and ROS levels were calculated as the ratio of fluorescence at 488 nm and 405 nm.

Oxygen consumption rates (OCR). For monolayer respiration assays, H460 cells were plated in growth media at 3×10^4 cells per well in XF24 microplates (Seahorse Bioscience; Billerica, MA) 24 h before the assay following manufacturer's recommendations for cell seeding. Growth media was changed to XF assay media and the plates were incubated at 37 °C in a non-CO_2 incubator for 45 min before starting the assay. To normalize the data, the cells were trypsinized and counted by haemocytometer or ViCELL automated cell counter (Beckman Coulter). To measure respiration in spheroids, H460 Spheroids were grown for 3 days in 96-well Ultra-Low Attachment round bottom plates (Sigma; Cat# CLS7007) starting with 2×10^3 cells per well, resulting in an average spheroid diameter of 401 ± 13 μm. For the respiration assay, spheroids were transferred to a Poly-D-lysine-treated XFe96 Spheroid Microplate (Seahorse Bioscience cat# 102959-100) containing 37 °C XF assay media at pH 7.4. Following each assay, the spheroid diameter was measured using images acquired on a Cytation 3 Cell Imaging Reader (BioTek; Winooski, VT). Volume was then calculated using the equation $v = {}^4/_3\pi r^3$. To normalize the data, the cell number of each spheroid was calculated by dividing the spheroid volume by the single-cell volume (average diameter of H460 cells = 14 μm). This method of calculating spheroid cell number was independently validated in parallel cultures by digesting spheroids with trypsin and manually counting the cells with a haemocytometer.

Determination of wild type IDH1 and IDH2 inhibition. IDH1 was generated as previously described[29]. IDH2 was expressed in Sf9 cells by baculovirus infection. Cells were lysed in lysis buffer (50 mM Tris, pH 7.5, 300 mM NaCl, 10% glycerol, 1% Triton) by Avestin emulsiflex C50. The supernatants were mixed with anti-Flag resin at 4 °C overnight and then eluted with addition of 100 μg ml^{-1} Flag peptide in lysis buffer (without Triton). The eluate was concentrated and loaded onto a Superdex 200 size exclusion column and eluted with sizing buffer (25 mM Tris, 100 mM NaCl, 1 mM DTT, pH 7.5). Fractions containing dimer were pooled and concentrated for use in kinetics. IDH activity was assessed by measuring the evolution of NADPH abundance using a coupled diaphorase/resazurin fluorescence assay[30]. Reactions were conducted at room temperature in 384-well Greiner black microtitre plates in a total volume of 10 μl of assay buffer. Final

compound concentrations were typically varied from 5 to 100,000 nM; the isocitrate concentration was fixed at 10 μM, the $NADP^+$ concentration was fixed at 5 μM, and IDH was fixed at 0.1 nM. Reactions were conducted in quadruplicate and run kinetically. After each addition, plates were centrifuged for 60 s to ensure complete mixing of reagents.

Data were fit to the following equation to determine the IC_{50}:

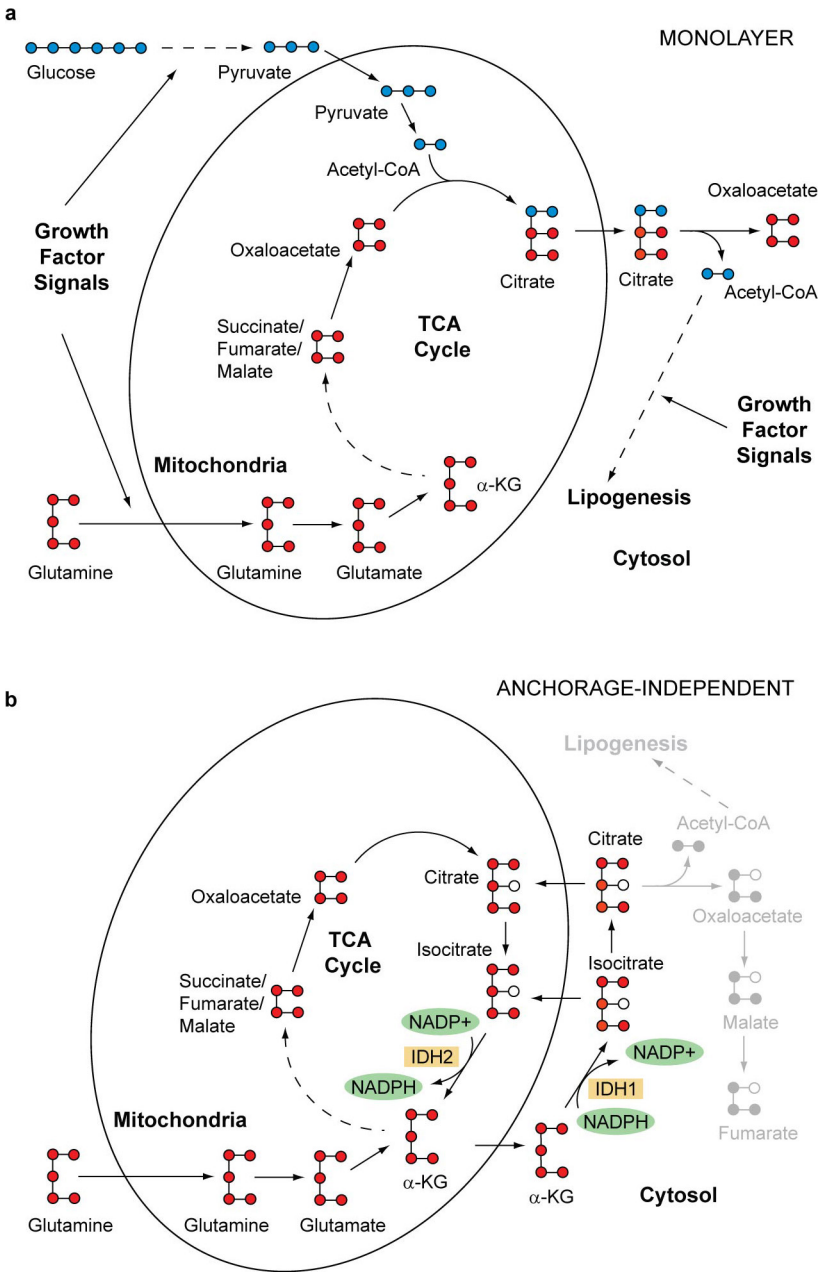
$$y = \frac{100\%}{1 + \left(\frac{x}{IC_{50}}\right)^s}$$

where y is the % of normalized enzyme activity, x is the concentration of inhibitor, and s is the Hill slope factor.

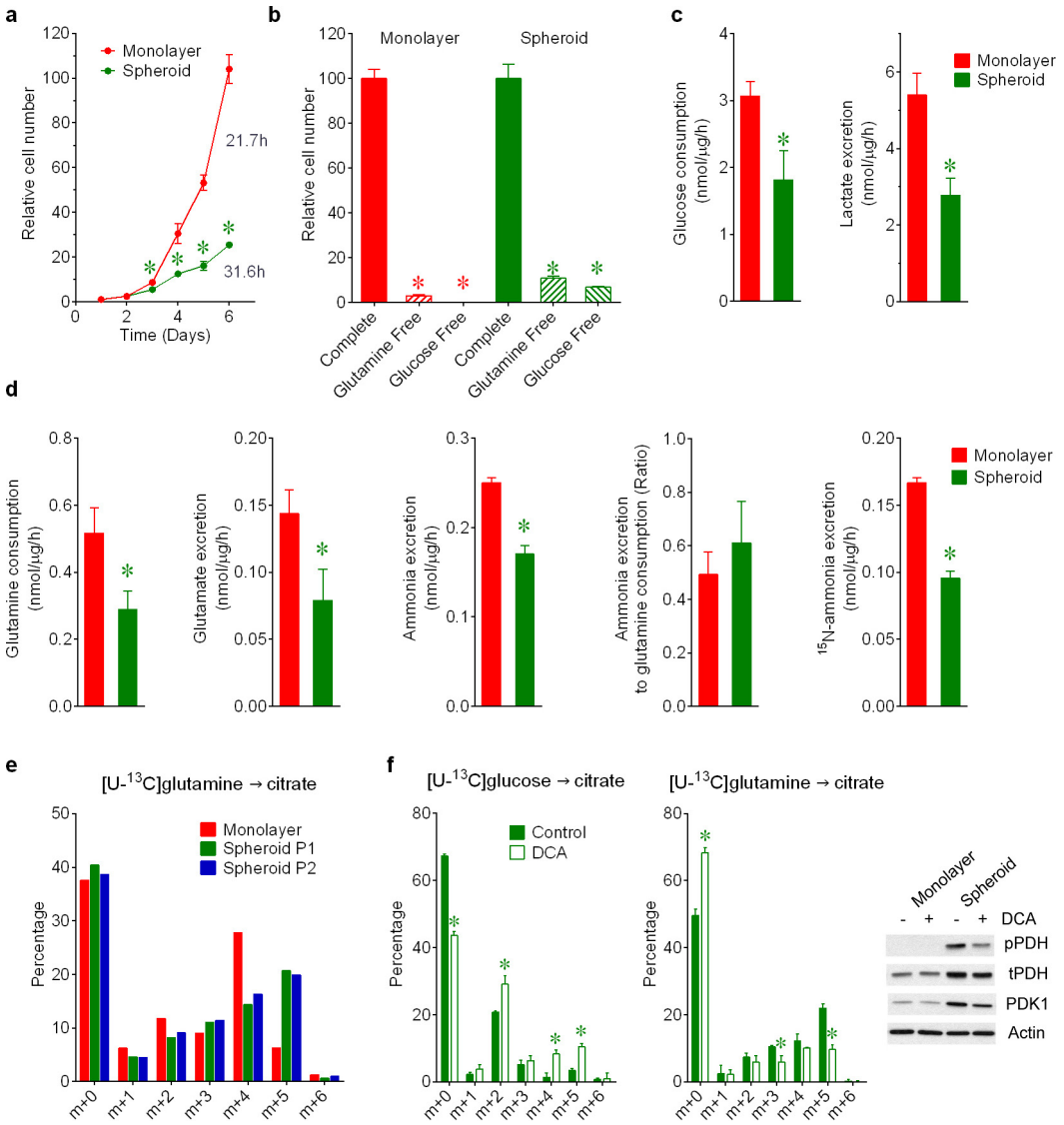
Statistics. No statistical methods were used to predetermine sample size. The experiments were not randomized, and the investigators were not blinded to allocation during experiments and outcome assessment.

Experiments in Figs 2a, 2b, 4f, and Extended Data Figs 5d, 7d and 8e were performed twice, and all other experiments were performed 3 times or more. Variation is always indicated using standard deviation. To assess the significance of differences among cultures and conditions, a two-tailed Welch's unequal variances t-test was used to assess the significance between two groups. For three or more groups, a one-way ANOVA followed by Dunnett's multiple comparisons test was performed. Before applying ANOVA, we first tested whether there was homogeneity of variation among the groups (as required for ANOVA) using the Brown–Forsythe test. Where the assumption of equal variance was violated, a $\log_2$ transformation was applied to the data before analysis. In a few cases, when significant differences in variation among groups persisted after transformation, we used Welch's unequal variances t-test followed by multiple-comparison correction.

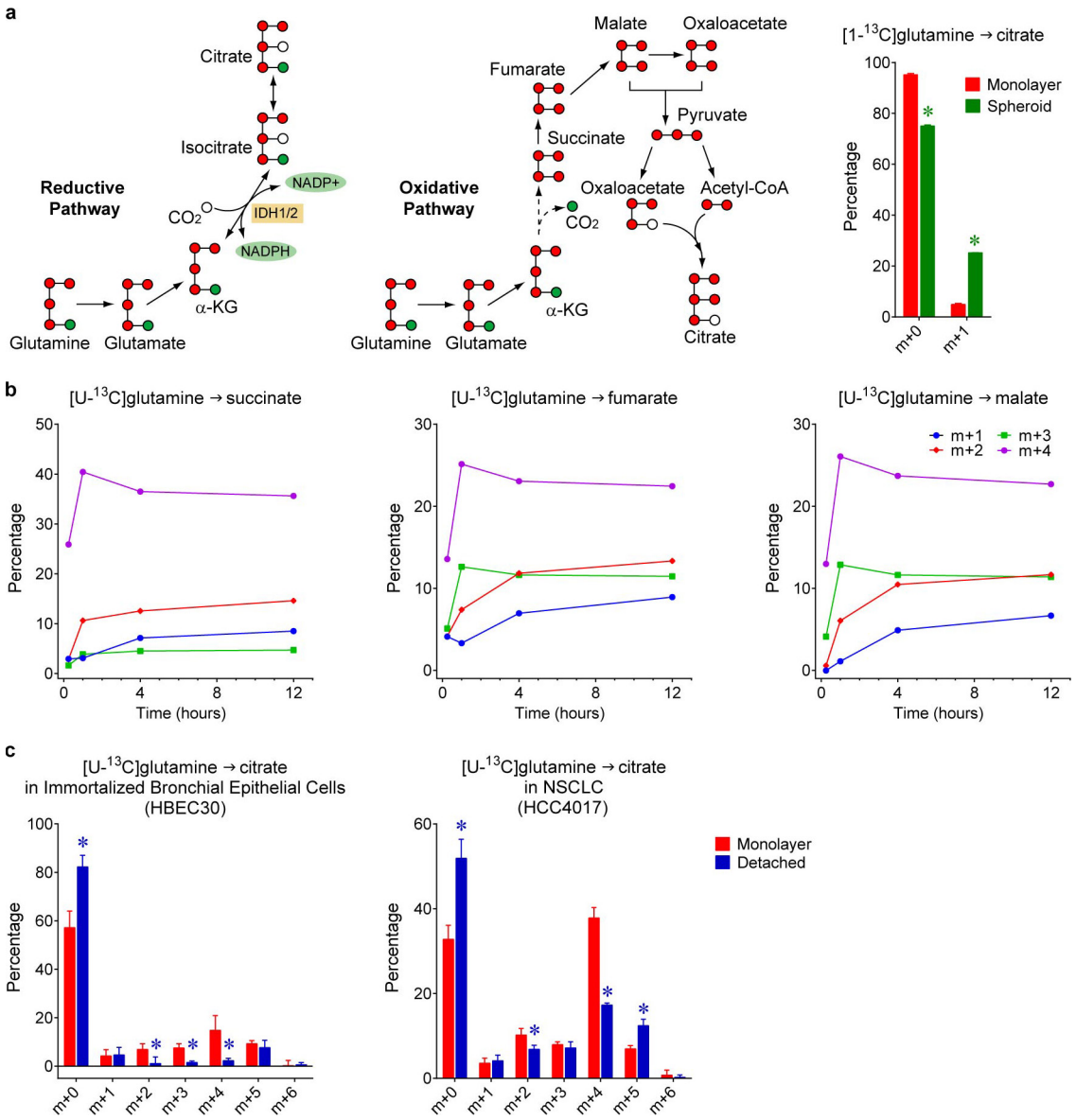
21. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nature Protocols* **8,** 2281–2308 (2013).
22. Yang, C. *et al.* Glutamine oxidation maintains the TCA cycle and cell survival during impaired mitochondrial pyruvate transport. *Mol. Cell* **56,** 414–424 (2014).
23. Cheng, T. *et al.* Pyruvate carboxylase is required for glutamine-independent growth of tumor cells. *Proc. Natl Acad. Sci. USA* **108,** 8674–8679 (2011).
24. Young, J. D. INCA: a computational platform for isotopically non-stationary metabolic flux analysis. *Bioinformatics* **30,** 1333–1335 (2014).
25. Antoniewicz, M. R., Kelleher, J. K. & Stephanopoulos, G. Elementary metabolite units (EMU): a novel framework for modeling isotopic distributions. *Metab. Eng.* **9,** 68–86 (2007).
26. Young, J. D., Walther, J. L., Antoniewicz, M. R., Yoo, H. & Stephanopoulos, G. An elementary metabolite unit (EMU) based method of isotopically nonstationary flux analysis. *Biotechnol. Bioeng.* **99,** 686–699 (2008).
27. Waleh, N. S. *et al.* Mapping of the vascular endothelial growth factor-producing hypoxic cells in multicellular tumor spheroids using a hypoxia-specific marker. *Cancer Res.* **55,** 6222–6226 (1995).
28. Wu, R. F., Ma, Z., Liu, Z. & Terada, L. S. Nox4-derived H_2O_2 mediates endoplasmic reticulum signaling through local Ras activation. *Mol. Cell. Biol.* **30,** 3553–3568 (2010).
29. Pietrak, B. *et al.* A tale of two subunits: how the neomorphic R132H IDH1 mutation enhances production of αHG. *Biochemistry* **50,** 4804–4812 (2011).
30. Rendina, A. R. *et al.* Mutant IDH1 enhances the production of 2-hydroxyglutarate due to its kinetic mechanism. *Biochemistry* **52,** 4563–4577 (2013).



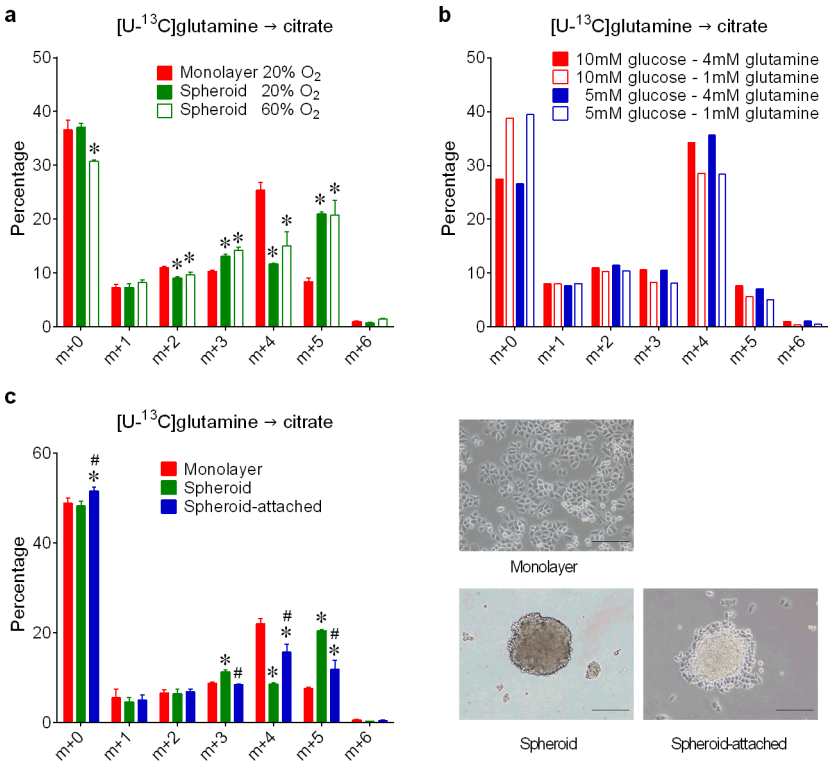
Extended Data Figure 1 | Alternative pathways of isocitrate/citrate metabolism. a, Predominant path of citrate formation in monolayer culture. **b**, Proposed pathway in anchorage-independent culture, emphasizing an alternative route of isocitrate/citrate metabolism and reducing equivalent flow.



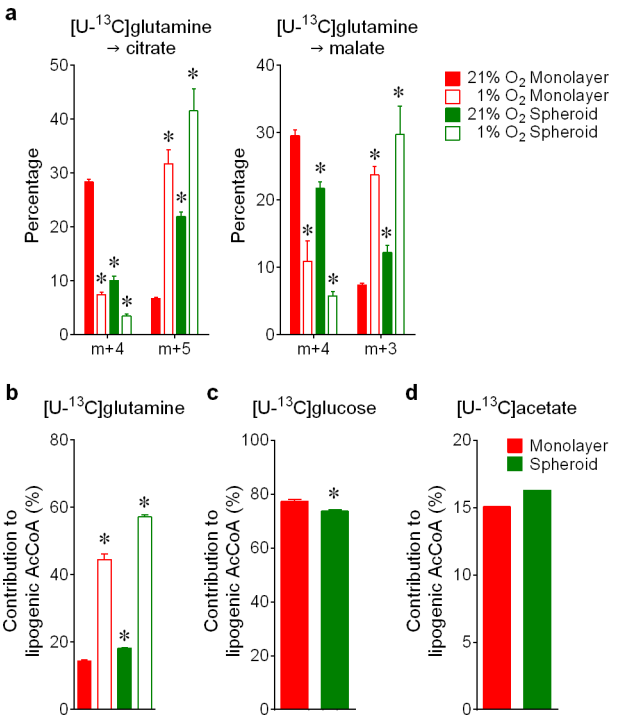
Extended Data Figure 2 | Nutrient metabolism in H460 spheroid culture. a, Cell proliferation and doubling times of H460 cells cultured under monolayer and spheroid conditions ($n = 4$ cultures days 1–4; $n = 3$ cultures days 5–6 from a representative experiment). **b**, Effect of glucose or glutamine deprivation on cell counts in monolayer and spheroid culture ($n = 4$ cultures from a representative experiment). **c**, Rates of glucose consumption and lactate excretion in monolayer and spheroid culture ($n = 4$ cultures from two experiments). **d**, Rates of glutamine consumption; glutamate and ammonia excretion; ratio of ammonia excretion to glutamine consumption; and rate of excretion of $^{15}NH_4^+$ originating from [γ-^{15}N]glutamine in monolayer and spheroid culture ($n = 3$ cultures from a representative experiment). **e**, Citrate mass isotopologue analysis in H460 cells in monolayer culture, aggregated into spheroids (P1), or disaggregated from spheroids then permitted to re-aggregate (P2) ($n = 2$ cultures from a representative experiment). **f**, Right, protein levels of phosphorylated PDHα (pPDH, Ser293), total PDHα (tPDH) and PDK1 in monolayer and spheroid culture with or without 2 mM dichloroacetate (DCA). Left, citrate mass isotopologue analysis in H460 spheroids cultured with [U-^{13}C]glucose or [U-^{13}C]glutamine, and treated with 2 mM DCA ($n = 3$ cultures from a representative experiment). All data represent mean ± s.d. *$P < 0.05$, Welch's unequal variances t-test (**a**, **c**, **d** and **f**), or Welch's unequal variances t-test, followed by multiple-comparison correction (**b**). All experiments were repeated 3 times or more.



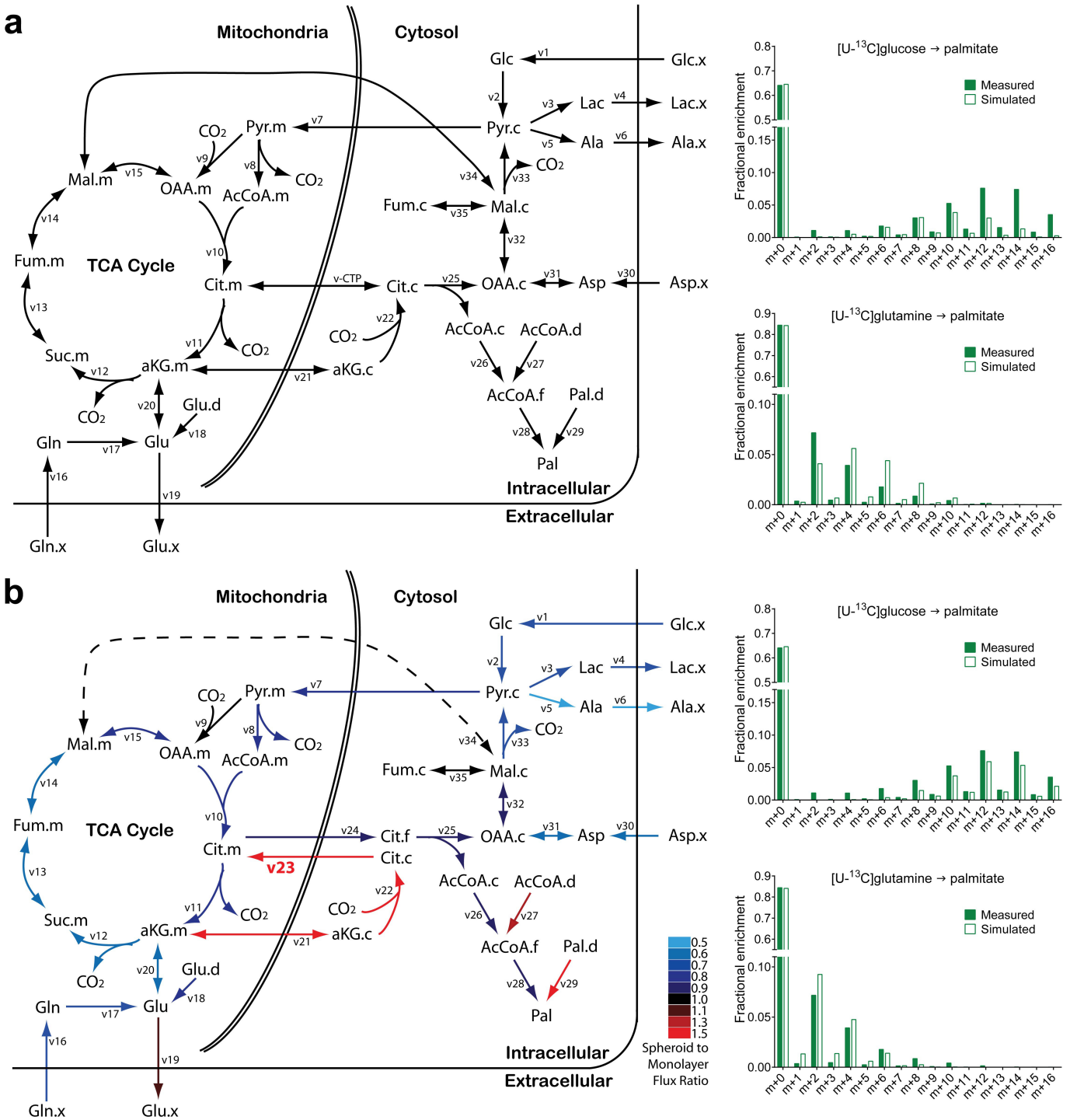
Extended Data Figure 3 | Reductive citrate metabolism in anchorage-independent spheroid culture. a, Citrate m+5 may be generated in several ways from [U-^{13}C]glutamine, including through reductive (left) or oxidative (middle) pathways. To test whether citrate m+5 arises from oxidative or reductive metabolism, spheroids were cultured with [1-^{13}C]glutamine. Glutamine-C1 (green circle) is released as CO_2 by α-ketoglutarate dehydrogenase in the oxidative TCA cycle, but is transferred to citrate via reductive metabolism. Citrate mass isotopologues in H460 cells cultured with [1-^{13}C]glutamine (right). The m+1 fraction in this experiment is comparable to the m+5 fraction from [U-^{13}C] glutamine (~20%), indicating that reductive labelling was enhanced in spheroids ($n = 3$ cultures from a representative experiment). **b**, Time-dependent evolution of succinate, fumarate and malate mass isotopologues in spheroids cultured with [U-^{13}C]glutamine ($n = 2$ cultures for each time point). **c**, Citrate labelling from [U-^{13}C]glutamine in immortalized, non-transformed bronchial epithelial cells (HBEC30) and lung cancer cells (HCC4017) from the same patient ($n = 3$ cultures from two experiments). All data represent mean ± s.d. *$P < 0.05$, Welch's unequal variances t-test. All experiments were repeated 3 times or more.



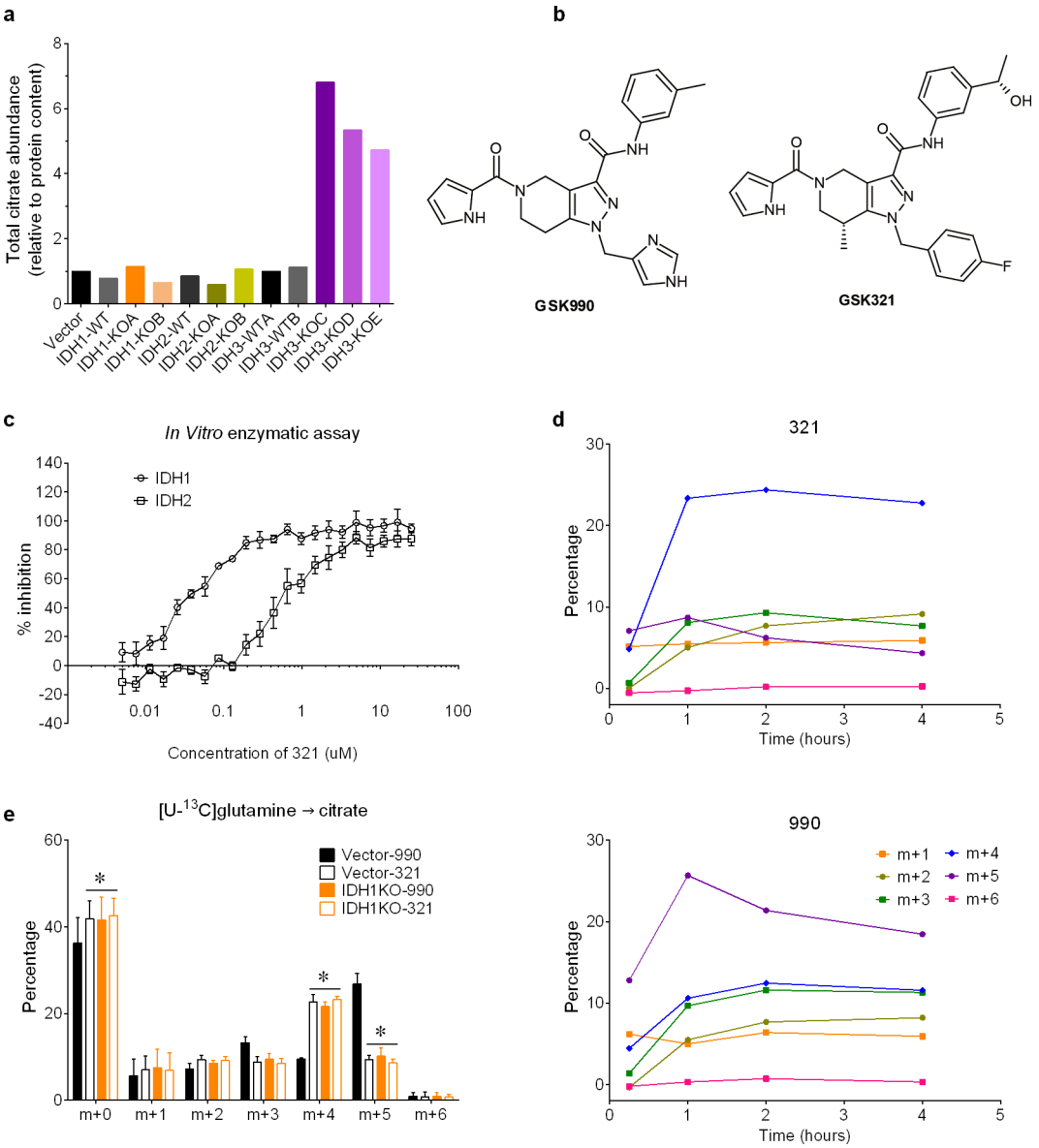
Extended Data Figure 4 | Effect of oxygen availability, nutrient availability and anchorage on reductive carboxylation. a, Mass isotopologues of citrate in spheroids cultured with [U-^{13}C]glutamine under 20% and 60% oxygen ($n = 4$ cultures from two experiments). **b**, Effects of reducing extracellular glucose and glutamine concentrations on citrate mass isotopologues in monolayer cells cultured with [U-^{13}C] glutamine ($n = 2$ cultures from a representative experiment). **c**, Day 7 spheroids were allowed to attach to a conventional tissue culture dish for 24 h, and mass isotopologues of citrate were analysed with [U-^{13}C] glutamine tracing ($n = 4$ cultures from two experiments). Insets are photomicrographs of cells in each of the culture conditions. Scale bars, 200 μm. All data represent mean ± s.d. $^{*}P < 0.05$ comparing to monolayer, $^{\#}P < 0.05$ comparing to spheroid, Welch's unequal variances *t*-test followed by multiple-comparison correction. All experiments were repeated 3 times or more.



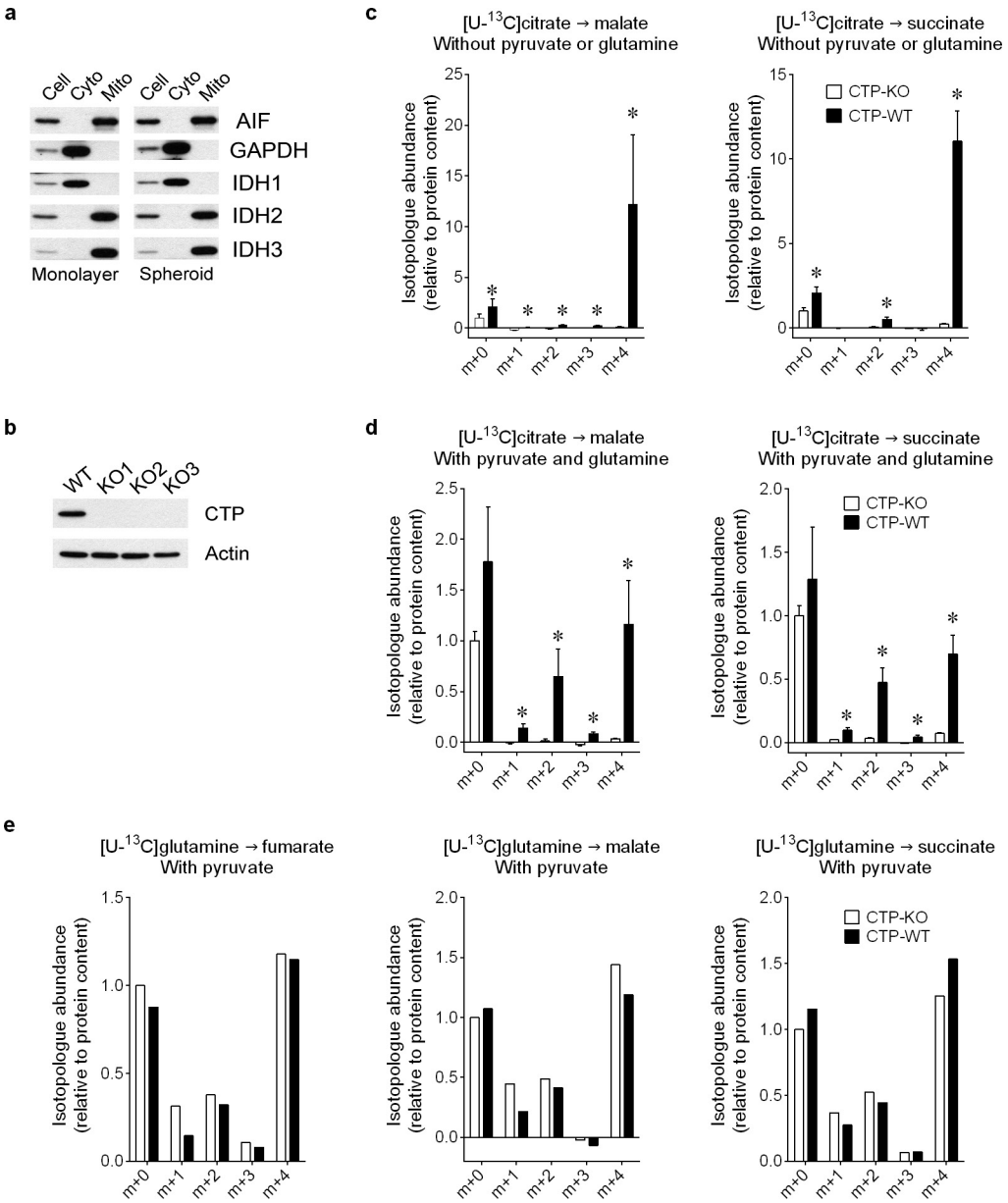
Extended Data Figure 5 | Hypoxia and anchorage independence elicit different effects on citrate metabolism. a, Mass isotopologues of citrate and malate in H460 cells cultured with [U-^{13}C]glutamine, in monolayer or spheroid conditions, under 21% or 1% oxygen ($n = 4$ cultures from two experiments). **b**–**d**, Contribution of glutamine (**b**), glucose (**c**) and acetate (**d**) to the lipogenic acetyl-CoA pool used for palmitate synthesis, during 24 h of culture with each tracer ($n = 3$ cultures in panels **b** and **c**; $n = 2$ cultures in panel **d** from a representative experiment). All data represent mean ± s.d. *$P < 0.05$, Welch's unequal variances t-test followed by multiple-comparison correction (**a**), or ANOVA (**b**), or Welch's unequal variances t-test (**c**). The experiment in **d** was repeated twice, and all other experiments were repeated 3 times or more.



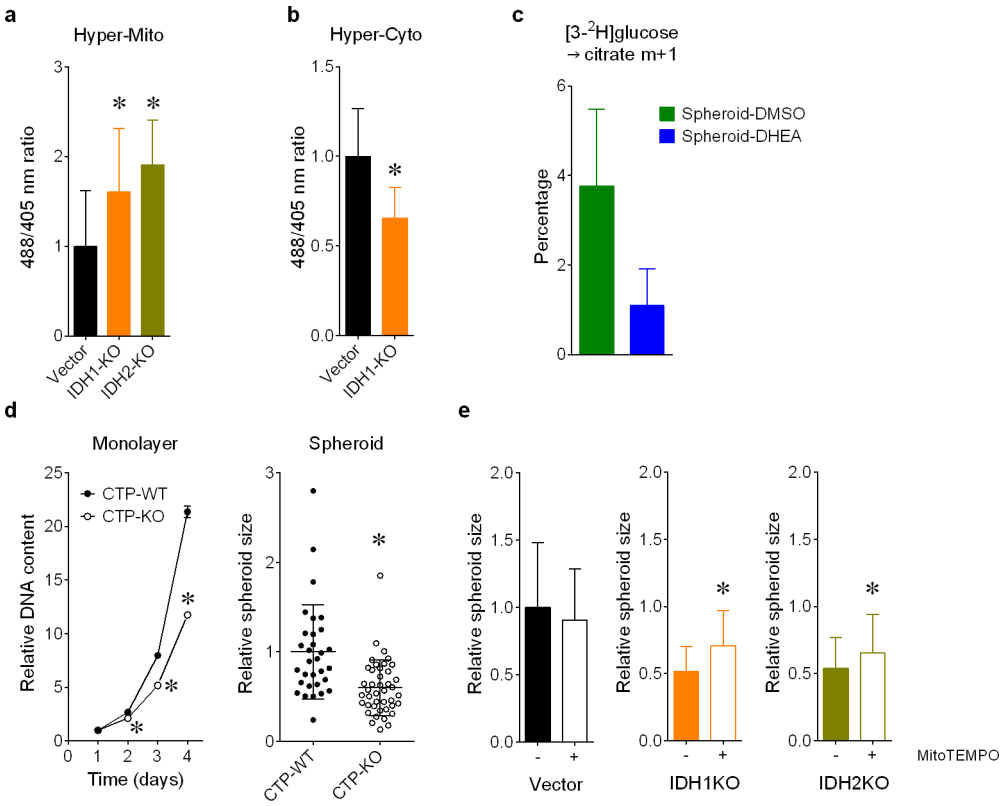
Extended Data Figure 6 | Graphical view of metabolic flux analysis (MFA). a, A conventional set of metabolic reactions and compartmentation produced an inadequate fit with the spheroid experimental data, with an unacceptable sum-of-squared residuals (SSR = 336). Poorly fit palmitate isotopologues from [U-^{13}C]glucose and [U-^{13}C]glutamine are shown on the right. v-CTP, bidirectional isocitrate/citrate trafficking flux. **b**, In the modified metabolic network, isocitrate/citrate produced from cytosolic reductive carboxylation enters the mitochondria and mixes with the isocitrate/citrate pool there. Adding this new reaction to the model (indicated in bold as v23) substantially improved the overall fit (SSR = 179) and the fit with palmitate isotopologues (right). Colour coding in **b** reflects flux changes in spheroids, expressed as the ratio of spheroid flux/monolayer flux. The dashed line indicates that the overall direction of malate transport was predicted to reverse from mitochondrial efflux in monolayer cells to mitochondrial import in spheroids. Flux terms are defined in Supplementary Table 2, and abbreviations and quantitative flux rates are in Extended Data Table 1.



Extended Data Figure 7 | IDH1 inhibition suppresses reductive carboxylation in spheroids. a, Abundance of citrate in vector control, wild-type, and IDH1, IDH2 or IDH3-deficient spheroids ($n = 2$ cultures from a representative experiment). **b**, Structure of IDH1 inhibitor compound GSK321 and structurally similar control compound GSK990. Compound GSK321 was initially identified as a potent inhibitor against an oncogenic allele of IDH1 containing the R132H mutation. Subsequent analysis revealed that at higher doses, the compound also inhibits wild-type IDH1. **c**, *In vitro* activity assay revealing effects of compound 321 on enzymatic activity of recombinant wild-type IDH1 and IDH2 ($n = 4$ repeats from a representative experiment). **d**, Time-dependent evolution of citrate mass isotopologues in 990- or 321-treated spheroids cultured with [U-^{13}C]glutamine ($n = 1$ culture for each time point). **e**, Mass isotopologues of citrate in vector control and IDH1KO spheroids cultured with [U-^{13}C]glutamine and treated with 5 μM IDH1 inhibitor (321) or control compound (990) ($n = 4$ cultures from two experiments). All data represent mean ± s.d. *$P < 0.05$, ANOVA. Experiments in **d** were repeated twice, and experiments in **c** and **e** were repeated 3 times or more.



Extended Data Figure 8 | Mitochondria take up and metabolize citrate. a, Protein expression of mitochondrial and cytosolic markers in subcellular fractions of monolayer and spheroid culture. **b**, Protein expression of CTP in control and CTP-deficient H460 cells. **c**, Mass isotopologues of malate and succinate in isolated mitochondria cultured with [U-^{13}C]citrate ($n = 4$ cultures from two experiments). **d**, Mass isotopologues of malate and succinate in isolated mitochondria cultured with [U-^{13}C]citrate, unlabelled pyruvate and glutamine ($n = 4$ cultures from two experiments). **e**, Mass isotopologues of fumarate, malate and succinate in isolated mitochondria cultured with [U-^{13}C]glutamine and unlabelled pyruvate ($n = 2$ cultures from a representative experiment). All data represent mean ± s.d. *$P < 0.05$, Welch's unequal variances t-test. Experiments in **e** were repeated twice, and all other experiments were repeated 3 times or more.



Extended Data Figure 9 | Reductive glutamine metabolism mitigates mitochondrial ROS and promotes spheroid growth. Mitochondrial (**a**) and cytosolic (**b**) ROS detected by a genetic hydrogen peroxide sensor in H460 spheroids containing or lacking IDH1 or IDH2. ($n = 29$ Vector spheroids; $n = 26$ IDH1-KO and IDH2-KO spheroids in panel **a**; $n = 23$ Vector spheroids; $n = 22$ IDH1-KO spheroids in panel **b**). **c**, Deuterium labelling of citrate in H460 spheroids without and with the pentose phosphate pathway inhibitor DHEA ($n = 3$ cultures from a representative experiment). **d**, Growth of H460 cells containing or lacking CTP in monolayer conditions (left) and as spheroids (right) ($n = 6$ monolayer cultures; $n = 31$ CTP-WT spheroids; $n = 42$ CTP-KO spheroids). **e**, Size of H460 spheroids containing or lacking IDH1 or IDH2, and treated with or without the mitochondrial ROS scavenger MitoTEMPO. ($n = 40$ Vector "-" spheroids; $n = 52$ Vector "+" spheroids; $n = 46$ IDH1KO "-" and "+" spheroids; $n = 48$ IDH2KO "-" spheroids; $n = 52$ IDH2KO "+" spheroids). All data represent mean ± s.d. $^*P < 0.05$, ANOVA (**a**), or Welch's unequal variances t-test (**b**–**e**). All experiments were repeated 3 times or more.

Extended Data Table 1 | Simulated metabolic fluxes in monolayer and spheroid culture

	Net Flux	Monolayer		Spheroid	
	Reaction	Value	95% C.I.	Value	95% C.I.
v1	Glc Uptake	1.0583	[1.0408,1.0759]	0.7453	[0.7281,.07626]
v2	Glycolysis	1.0583	[1.0408,1.0759]	0.7453	[0.7281,.07626]
v3	LDH	1.8558	[1.8366,1.8749]	1.2458	[1.2268,1.2649]
v4	Lac Secretion	1.8558	[1.8366,1.8749]	1.2458	[1.2268,1.2649]
v5	GPT	0.0896	[0.0857,0.0935]	0.0475	[0.0436.0.0514]
v6	Ala Secretion	0.0896	[0.0857,0.0935]	0.0475	[0.0436.0.0514]
v7	MPC	0.5763	[0.5361,0.6176]	0.4731	[0.4336,0.5130]
v8	PDH	0.5236	[0.4844,0.5639]	0.4200	[0.3813,0.4587]
v9	PC	0.0527	[0.0452,0.0603]	0.0532	[0.0473,0.0592]
v10	CS	0.5236	[0.4844,0.5639]	0.4200	[0.3813,0.4587]
v11	IDH.m	0.2510	[0.2371,0.2655]	0.2004	[0.1879,0.2146]
v12	OGDH	0.4824	[0.4611,0.5041]	0.2977	[0.2826,0.3118]
v13	SDH	0.4824	[0.4611,0.5041]	0.2977	[0.2826,0.3118]
v14	FH.m	0.4824	[0.4611,0.5041]	0.2977	[0.2826,0.3118]
v15	MDH.m	0.4709	[0.4312,0.5115]	0.3668	[0.3280,0.4052]
v16	Gln Uptake	0.2615	[0.2537,0.2692]	0.1731	[0.1692,0.1770]
v17	GLS	0.2615	[0.2537,0.2692]	0.1731	[0.1692,0.1770]
v18	Glu Dilution	0.0722	[0.0632,0.0814]	0.0592	[0.0540,0.0644]
v19	Glu Secretion	0.0393	[0.0354,0.0432]	0.0423	[0.0384,0.0462]
v20	GDH	0.2943	[0.2808,0.3081]	0.1900	[0.1819,0.1982]
v21	aKG Transporter	0.0629	[0.0547,0.0715]	0.0926	[0.0848,0.1068]
v22	IDH.c	0.0629	[0.0547,0.0715]	0.0926	[0.0848,0.1068]
v23	CTP Influx	0.0629	[0.0547,0.0715]	0.0926	[0.0848,0.1068]
v24	CTP Efflux	0.3355	[0.2929,0.3790]	0.3122	[0.2699,0.3546]
v25	ACLY	0.3355	[0.2929,0.3790]	0.3122	[0.2699,0.3546]
v26	AcCoA Tracer	0.3355	[0.2929,0.3790]	0.3122	[0.2699,0.3546]
v27	AcCoA Dilution	0.0121	[0.0021,0.0234]	0.0159	[0.0000,0.0371]
v28	FASN	0.0434	[0.0377,0.0493]	0.0410	[0.0348,0.0475]
v29	Pal Dilution	0.0638	[0.0542,0.0746]	0.1161	[0.0968,0.1396]
v30	Asp Uptake	0.0581	[0.0469,0.0706]	0.0326	[0.0265,0.0415]
v31	AST.c	0.0581	[0.0469,0.0706]	0.0326	[0.0265,0.0415]
v32	MDH.c	0.3936	[0.3501,0.4385]	0.3448	[0.3012,0.3993]
v33	ME.c	0.4051	[0.3882,0.4226]	0.2758	[0.2644,0.2883]
v34	Mal Transporter	-0.0115	[-0.0538,0.0322]	0.0690	[0.0271,0.1109]
v35	FH.c	0.0000	[0.0000,0.0000]	0.0000	[0.0000,0.0000]

	Exchanging	Monolayer		Spheroid	
	Reaction	Value	95% C.I.	Value	95% C.I.
v13.e	SDH	17.7187	[0.0000,inf]	4.4558	[0.0000,inf]
v14.e	FH.m	15841	[0.0000,inf]	15020	[1.2075,inf]
v15.e	MDH.m	0.8754	[0.1993,6.9294]	76030	[1.2040,inf]
v20.e	GDH	1.6515	[1.2667,2.2882]	1.3421	[1.0074,inf]
v21.e	aKG Transporter	5.0559	[0.0000,inf]	12.8126	[0.0000,inf]
v31.e	AST.c	0.5063	[0.2754,1.7111]	0.1392	[0.1009,0.2211]
v32.e	MDH.c	0.9824	[0.1811,inf]	0.0000	[0.0000,inf]
v34.e	Mal Transporter	0.7175	[0.3974,1.3969]	0.0000	[0.0000,0.0621]
v35.e	FH.c	28139	[0.9790,inf]	102230	[3.4108,inf]

	Mixing	Monolayer		Spheroid	
	Reaction	Value	95% C.I.	Value	95% C.I.
v36	Cit.cyto	0.0000	[0.0000,0.0080]	0.4688	[0.4355,0.4981]
v37	Cit.mito	1.0000	[0.9920,1.0000]	0.5312	[0.5019,0.5645]
v38	Cit.mix	1.0000	[1.0000,1.0000]	1.0000	[1.0000,1.0000]
v39	Mal.cyto	1.0000	[0.7622,1.0000]	1.0000	[0.8410,1.0000]
v40	Mal.mito	0.0000	[0.0000,0.2378]	0.0000	[0.0000,0.1590]
v41	Mal.mix	1.0000	[1.0000,1.0000]	1.0000	[1.0000,1.0000]
v42	Fum.cyto	0.9902	[0.6662,1.0000]	0.9918	[0.8278,1.0000]
v43	Fum.mito	0.0098	[0.0000,0.3338]	0.0082	[0.0000,0.1722]
v44	Fum.mix	1.0000	[1.0000,1.0000]	1.0000	[1.0000,1.0000]

Fluxes were determined using the network illustrated in Extended Data Fig. 6b. Negative flux values indicate that the net direction is reverse to the direction indicated in Supplementary Table 2. The sum of squared residuals was 168.8 for monolayer and 179 for spheroid, both within the acceptable range (117.1–184.7). Degrees of freedom = 149. Units for all fluxes are nmol h^{-1} per microgram protein.

Metabolites: AcCoA, acetyl-CoA; αKG, α-ketoglutarate; Ala, alanine; Asp, aspartate; Cit, citrate; Fum, fumarate; Glc, glucose; Gln, glutamine; Glu, glutamate; Lac, lactate; Mal, malate; Pal, palmitate. c, cytosolic; m, mitochondrial. C.I., Confidence Interval.

Enzymes and transporters: LDH, lactate dehydrogenase; GPT, glutamate pyruvate transaminase; MPC, mitochondrial pyruvate carrier; PDH, pyruvate dehydrogenase; PC, pyruvate carboxylase; CS, citrate synthase; IDH, isocitrate dehydrogenase; OGDH, α-ketoglutarate dehydrogenase; SDH, succinate dehydrogenase; FH, fumarase; MDH, malate dehydrogenase; GLS, glutaminase; GDH, glutamate dehydrogenase; CTP, citrate transporter protein; ACLY, ATP citrate lyase; FASN, fatty acid synthase; AST, aspartate aminotransferase; ME, malic enzyme.

LETTER

doi:10.1038/nature17437

Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes

Dilmi Perera[1]*, Rebecca C. Poulos[1]*, Anushi Shah[1], Dominik Beck[1], John E. Pimanda[1,2] & Jason W. H. Wong[1]

Promoters are DNA sequences that have an essential role in controlling gene expression. While recent whole cancer genome analyses have identified numerous hotspots of somatic point mutations within promoters, many have not yet been shown to perturb gene expression or drive cancer development[1–4]. As such, positive selection alone may not adequately explain the frequency of promoter point mutations in cancer genomes. Here we show that increased mutation density at gene promoters can be linked to promoter activity and differential nucleotide excision repair (NER). By analysing 1,161 human cancer genomes across 14 cancer types, we find evidence for increased local density of somatic point mutations within the centres of DNase I-hypersensitive sites (DHSs) in gene promoters. Mutated DHSs were strongly associated with transcription initiation activity, in which active promoters but not enhancers of equal DNase I hypersensitivity were most mutated relative to their flanking regions. Notably, analysis of genome-wide maps of NER[5] shows that NER is impaired within the DHS centre of active gene promoters, while *XPC*-deficient skin cancers do not show increased promoter mutation density, pinpointing differential NER as the underlying cause of these mutation hotspots. Consistent with this finding, we observe that melanomas with an ultraviolet-induced DNA damage mutation signature show greatest enrichment of promoter mutations, whereas cancers that are not highly dependent on NER, such as colon cancer, show no sign of such enrichment. Taken together, our analysis has uncovered the presence of a previously unknown mechanism linking transcription initiation and NER as a major contributor of somatic point mutation hotspots at active gene promoters in cancer genomes.

To examine systematically the frequency of somatic point mutations at gene promoters, we curated mutation calls from 1,161 whole cancer genomes across 14 cancer types from The Cancer Genome Atlas (TCGA), International Cancer Genome Consortium (ICGC) and published mutations[6] (Fig. 1a, see Supplementary Table 1 for full list of samples and sources). Using cell-type-specific epigenomic data, we calculated the average mutation density across promoter DHSs, enhancer DHSs, genic and heterochromatin regions (Methods). In many cancer types, the average mutation density at promoter DHSs was higher than that of enhancer DHSs, with melanoma, lung and

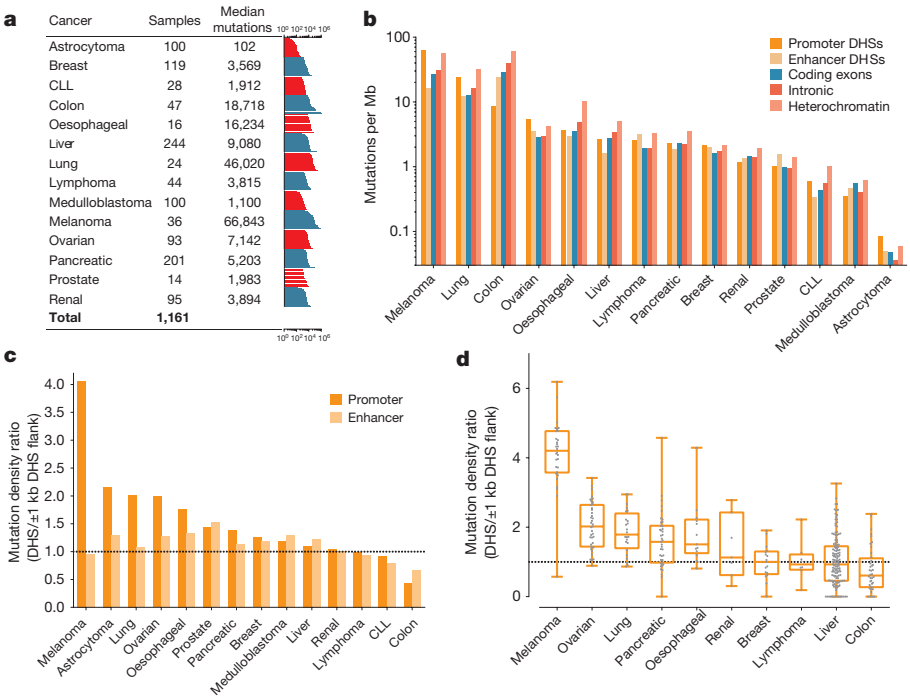


Figure 1 | Pan-cancer analysis of mutation distribution across the genome reveals enrichment at promoter DHSs. a, Summary of the number of samples and median mutations associated with each cancer type analysed, with a histogram for each cancer type showing the total mutation count across each sample. CLL, chronic lymphocytic leukaemia. **b**, Mutation density across selected genomic regions for each cancer type. DHSs are defined as ±75 bp of the DHS centre. **c**, Comparison of mutation density within promoter DHSs and enhancer DHSs relative to their flanking regions (±1 kb). **d**, Distribution of promoter DHS/DHS flank ratios of individual cancer samples with at least 8,602 mutations separated by cancer type. Box-plot shows median, quartiles, maximum and minimum. Individual cancer samples are shown as grey dots.

[1]Prince of Wales Clinical School and Lowy Cancer Research Centre, UNSW Australia, Sydney 2052, Australia. [2]Department of Haematology, Prince of Wales Hospital, Sydney 2031, Australia.
*These authors contributed equally to this work.

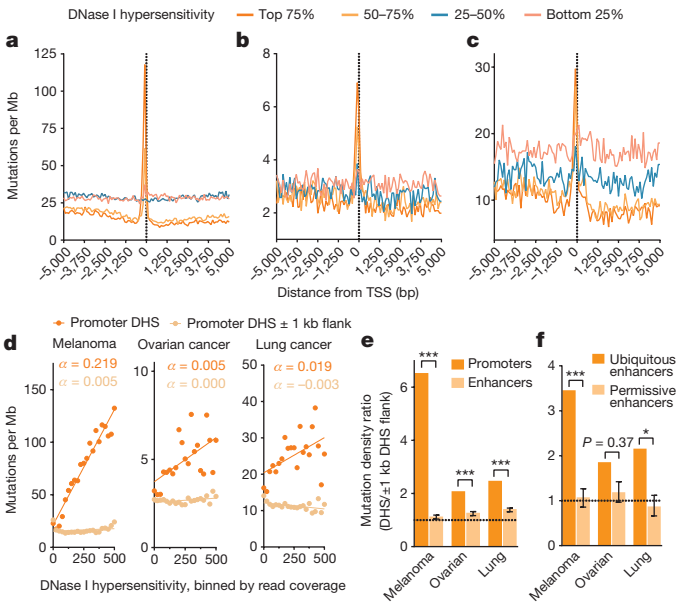


Figure 2 | Promoter mutation density is strongly linked with transcription activity. a–c, Mutation density profiles ±5 kb of TSSs for (**a**) melanoma, (**b**) ovarian cancer and (**c**) lung cancer stratified by the DNase I hypersensitivity of promoters as measured by DNase-seq read coverage of melanocytes (**a**), ovary cells (**b**) and A549 cells (**c**). **d**, Mutation density associated with DNase I hypersensitivity as determined by DNase-seq read coverage at promoter DHSs and promoter DHS flank ±1 kb in melanoma, ovarian cancer and lung cancer. The slope (α) was calculated from the linear regression on the binned values for mutation density versus DNase I hypersensitivity. **e**, Mutation density ratio of DHS centre/±1 kb DHS flank of DNase I hypersensitivity matched active promoters and enhancers, where active promoters are defined as those within the top 25% of DNase-seq coverage. The error bar on the enhancer data set shows the interquartile range of repair ratios over 100 randomized samplings of matched enhancers. **f**, Ubiquitous and permissive enhancers relative to their DHS flanking regions in melanoma, ovarian cancer and lung cancer. The full set of ubiquitous enhancers from FANTOM5 ($n = 200$) was used and a matching set of equally nucleosome free permissive enhancers was sampled and repeated 100 times as described above (also see Methods). $^*P < 0.05$, $^{***}P < 0.001$ (χ^2 test).

ovarian cancers being significantly higher at >1.5 times the mutation density (Fig. 1b, see Extended Data Table 1 for list of *P* values from χ^2 tests). Mutations in genic regions are generally lower as a result of transcription-coupled repair[7] and domain-associated repair[8], while heterochromatin DNA is typically most highly mutated owing to the inaccessibility of the DNA to repair enzymes[8–10].

To assess whether the mutation density is increased across the entire promoter region or is confined to the core nucleosome-free DHS centres, we examined the ratio of mutation density within promoter and enhancer DHSs (150 base pairs (bp)) relative to their flanking regions (± 1 kb of DHS). Notably, in melanoma, the mutation density at promoter DHS centres was more than four times that of their flanking regions (Fig. 1c, $P < 0.0001$, χ^2 test). Other cancers that showed a significant increase (>1.5-fold) in promoter DHS mutation density compared with their flanking regions include astrocytoma, lung, ovarian, oesophageal and prostate cancers (Fig. 1c, see Extended Data Table 2 for list of *P* values from χ^2 tests). Although gene promoters are typically GC-rich, adjusting for sequence GC content or trinucleotide frequencies did not significantly affect our observation (Extended Data Fig. 1a, Methods). Furthermore, mutation classes within the promoter DHS are broadly similar to those in their flanking regions (Extended Data Fig. 2). While mutation strand bias can be observed in the genic regions of melanoma and lung cancer, no strand bias was observed in the promoter DHS (Extended Data Fig. 1c–e). In contrast to the promoter DHS, on average, the enhancer DHS did not show substantial enrichment in mutation density (promoter DHS: range 4.06 to 0.43; enhancer DHS: range 1.52 to 0.67, Fig. 1c).

To determine whether the increased mutation density at promoter DHSs varies across individual cancer samples, we calculated the promoter DHS/DHS flank mutation density ratio for samples with >8,602 somatic point mutations. This threshold was chosen as it enables robust estimates of promoter DHS/DHS flank mutation density ratios of >2 (Extended Data Fig. 3a, Methods). Promoter DHS density relative to flanking regions varied between cancer samples (Fig. 1d), but was particularly prevalent in some cancers such as melanoma, ovarian and lung cancers (Fig. 1d). We did not find a single distinct somatic trinucleotide mutation signature across all cancer types associated with samples with increased promoter mutations (Extended Data Fig. 1b), suggesting that this phenomenon is not induced by a particular mutagen, but instead is more likely the result of an intrinsic underlying mechanism. Across all cancer samples, the promoter DHS/DHS flank mutation density ratio does not seem to be associated with genome-wide mutation density (Extended Data Fig. 3b), and the promoter mutations are generally dispersed across the genome (Extended Data Fig. 3c, d).

To identify the underlying cause of this increased promoter DHS mutation density, we applied logistic regression to model the influence of a variety of factors on the likelihood of a promoter being mutated. These factors include DNase I hypersensitivity, associated gene expression, replication timing, proportion of rare single nucleotide polymorphisms (SNPs), sequence conservation, whether the associated gene is a known cancer gene and potential sequence-dependent factors including the percentage GC content and trinucleotide frequencies within each promoter (Methods). These factors were selected as they have all been previously associated with variations in regional somatic mutation density[10–16]. Regression models were computed for melanoma, ovarian and lung cancers as only these cancers had sufficient mutation counts and showed enrichment of mutations in promoter DHSs. DNase I hypersensitivity was found to be most significantly associated with mutated gene promoters across all three cancer types (Extended Data Table 3). This association remained significant across all three cancer types, even after adjustment for multiple factors in a multivariate logistic regression model (Extended Data Table 3 and Supplementary Tables 6–8).

Examination of the mutation density profiles across the transcription start site (TSS) revealed a sharp increase in mutation density in the ~100 bp window upstream of the TSS in comparison to up- and downstream flanking regions with the strongest effect seen in the most DNase I-hypersensitive promoters (Fig. 2a–c). Similarly, we observed an increase in mutation density upstream of the TSS in highly expressed, but not lowly/non-expressed genes, suggesting a positive relationship between promoter activity and mutation density (Extended Data Fig. 4a–c). There was strong and significant positive correlation between promoter DHS mutation density and DNase I hypersensitivity ($P < 0.001$, Poisson regression; Supplementary Table 9 and Fig. 2d). This association with chromatin accessibility was generally present regardless of mutation class (Extended Data Fig. 5). The positive correlation between mutations and chromatin accessibility is in contrast to recent studies that found that over broader DHSs, such a correlation was negative[13,17]. Indeed, there is a significant negative association between the DNase I hypersensitivity and the mutation density of the ±1 kb region flanking the DHS in melanoma and lung cancer ($P < 0.001$, Poisson regression; Supplementary Table 9 and Fig. 2d). This suggests that the enrichment of point mutations at the DHS centre of active promoters is a phenomenon independent of the decreased mutation density over broad DHSs (see Supplementary Data for detailed comparative analysis with refs 13 and 15).

Analysis of mutation density thus far suggests that localized increased mutation density is present at promoter DHSs. In particular, mutation density was found to be highest within transcription factor binding footprints (Extended Data Fig. 4d, e, Methods). As

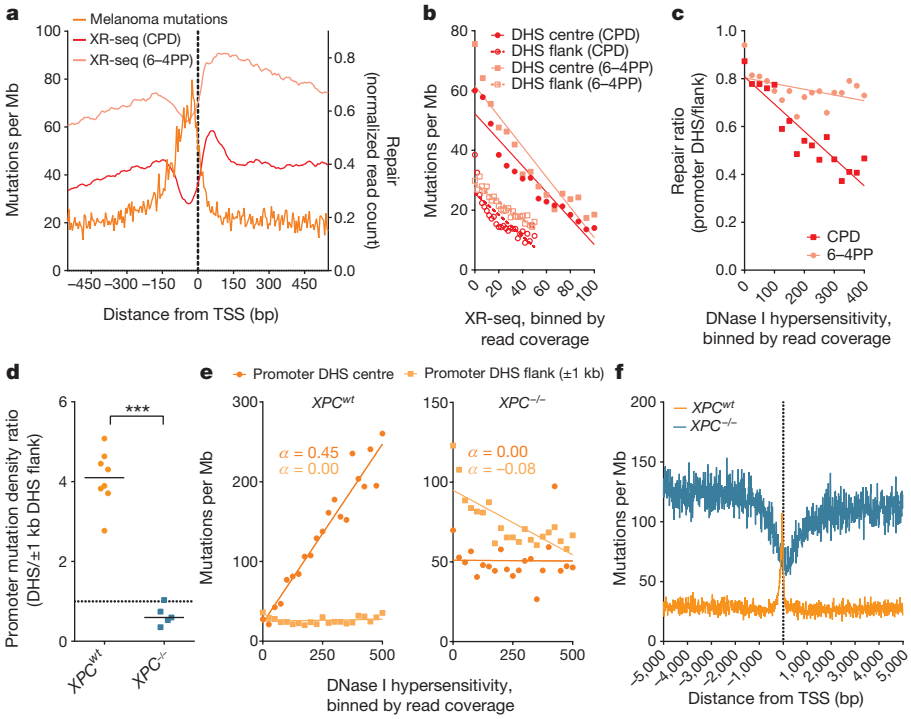


Figure 3 | Nucleotide excision repair in ultraviolet-irradiated human cells inversely mirrors mutation density in promoters and ubiquitous enhancers. a, Average melanoma mutation and XR-seq profiles for CPD and 6–4PP in normal skin fibroblast. **b**, Ratio of mutations per megabase (Mb) in melanoma associated with CPD and 6–4PP repair in promoter DHS centre (±75 bp) and DHS flank (±1 kb). A trend line is fitted by linear regression over the binned data for mutations per Mb versus repair density ratio. **c**, Ratio of repair of CPD and 6–4PP in promoter DHSs against the flanking region associated with DNase I hypersensitivity. A trend line is fitted by linear regression over the binned data. The flanking region is defined as ±150 bp of the DHS. **d**, Promoter mutation density ratio in XPC^{wt} ($n = 8$) and $XPC^{-/-}$ (NER deficient) ($n = 5$) SCC genomes. ***$P < 0.001$ (unpaired t-test). **e**, Mutation density associated with DNase I hypersensitivity at promoter DHSs and ±1 kb flank regions in XPC^{wt} and $XPC^{-/-}$ SCC. A trend line is fitted by linear regression on the binned values. **f**, Average mutation density profiles ±5 kb of all TSSs for XPC^{wt} and $XPC^{-/-}$ SCC. The slope (α) was calculated from the linear regression on the binned values for mutation density versus DNase I hypersensitivity.

transcription factor binding also occurs at enhancers, we sought to determine whether it is transcription initiation that is necessary for the increased mutation density at DHSs. To normalize for transcription factor binding, a set of the top 25% of DNase I hypersensitive promoters was compared with a set of enhancers with matched hypersensitivity (Methods). These enhancers did not show increased DHS mutation density >1.5-fold over flanking regions and the mutation density ratio was significantly lower than that of the promoter DHS ($P < 0.001$, χ^2 test, Fig. 2e). This suggests that transcription factor binding alone does not contribute to increased mutation density. Recent studies have shown that enhancers share many similar gene regulatory mechanisms as promoters, but that the strength of transcription initiation at enhancers is generally weaker than at promoters[18–20]. To test further whether transcription initiation activity underlies increased DHS mutation density, we compared a set of ubiquitous enhancers (which promote strongly the transcription of enhancer RNAs) with a set of permissive enhancers, both as defined by FANTOM5 (ref. 18), with matching DNase I hypersensitivity (Methods). Like active promoters, ubiquitous enhancers showed increased (>1.5-fold) enhancer DHS mutation density over flanking regions, and this was significantly higher than the same ratio in permissive enhancers, in both melanoma ($P < 0.001$, χ^2 test) and lung cancer ($P = 0.041$, χ^2 test) (Fig. 2f). Thus, mutation hotspots are not only present in promoters, but are also more generally linked with transcription initiation.

Differential NER[8,13] and mismatch repair[16] have been found to contribute to mutation density variation across cancer genomes at mega- to kilobase resolution. As there is also evidence that transcription factor binding can interfere with NER at specific gene promoters[21–23], we used genome-wide maps of NER (XR-seq[5]) to test whether differential NER at gene promoters accounts for the localized increase in mutation density. We generated cyclobutane pyrimidine dimer (CPD) and pyrimidine–pyrimidone (6–4) photoproduct (6–4PP) XR-seq profiles from normal skin fibroblasts across TSSs and compared these with mutation density profiles from melanoma (Methods). Notably, there is a clear depletion of NER ~100 bp upstream of the TSS, corresponding inversely with the increased mutation density in melanoma (Fig. 3a). NER for both types of lesions was significantly negatively correlated with mutation density both within the DHS centre and its flanking region ($P < 0.001$, Gaussian linear regression, Fig. 3b and Supplementary Table 9). Furthermore, the rate of repair 100 bp upstream of the TSS relative to the 150 bp flanking region decreased with increased DNase I hypersensitivity (Fig. 3c). Finally, as there was significant localized decreased CPD NER in promoters and ubiquitous enhancers, but not permissive enhancers (Extended Data Fig. 6a, b; $P < 0.001$, χ^2 test), this suggests that NER is specifically impaired by transcription initiation.

Xeroderma pigmentosum is a rare genetic disorder arising from inactivating germline mutations of enzymes involved in NER. Comparison of somatic point mutations in skin squamous cell carcinoma (SCC) genomes of patients with xeroderma pigmentosum containing germline *XPC* inactivation ($XPC^{-/-}$) and of non-xeroderma-pigmentosum patients ($XPC^{wildtype}$, hereafter described as XPC^{wt})[8] revealed that XPC^{wt} SCC genomes had a high promoter DHS/DHS flank ratio (median 4.1; Fig. 3d), and this ratio was significantly higher than $XPC^{-/-}$ where there was no increase in promoter DHS mutation density ($P < 0.001$, unpaired t-test, median 0.6, Fig. 3d). In comparison to XPC^{wt}, promoter mutations in $XPC^{-/-}$ SCC genomes did not show a positive association with DNase I hypersensitivity (Fig. 3e). Furthermore, increased mutation density was only present in promoters and ubiquitous enhancers of XPC^{wt} but not $XPC^{-/-}$ SCC genomes (Extended Data Fig. 6c, d). Notably, mutation density in $XPC^{-/-}$ SCCs decreased across the TSS (Fig. 3f). This may

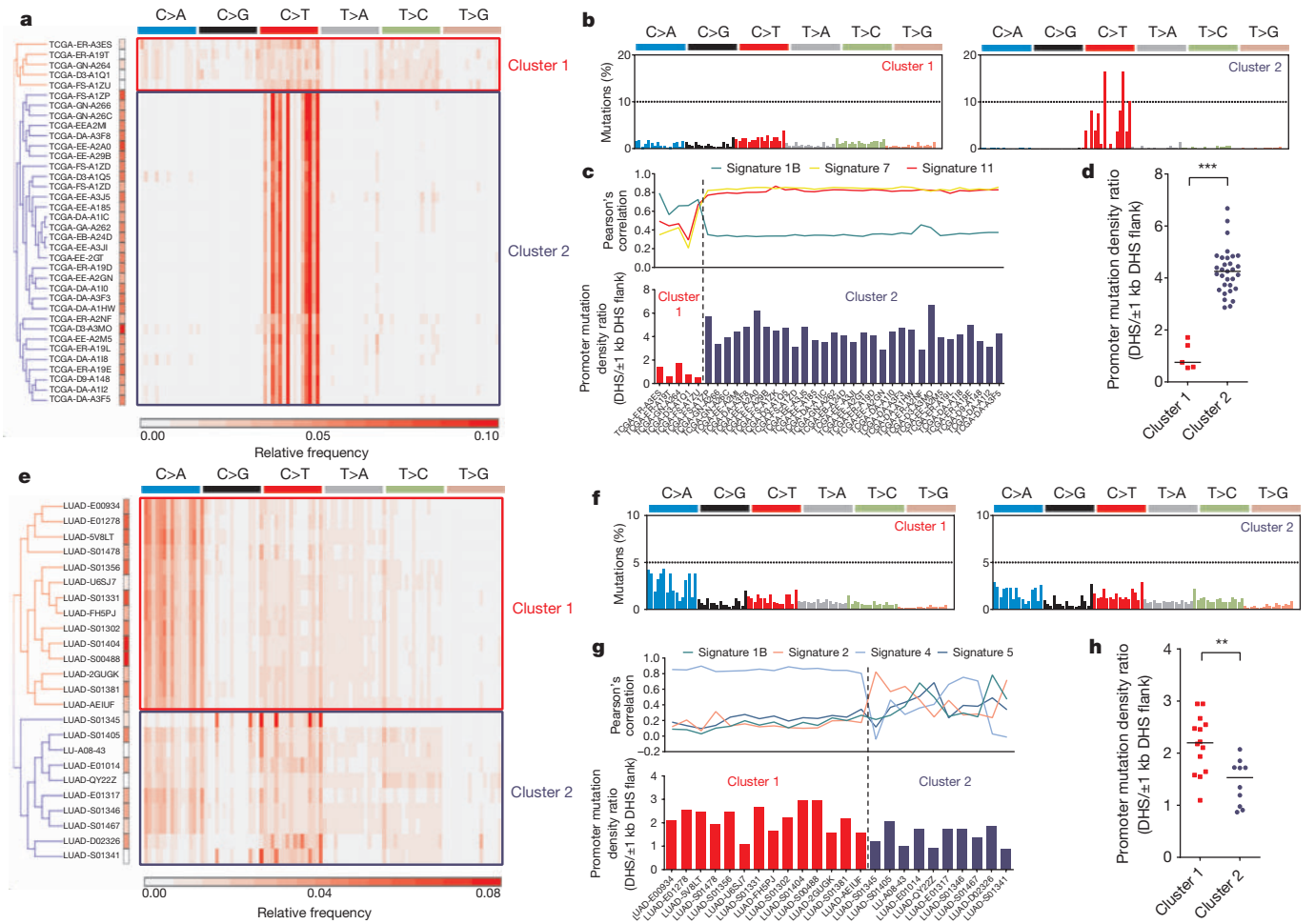


Figure 4 | Mutation signatures and promoter DHS mutation density in melanoma and lung cancer. a, **e**, Heatmaps showing the relative frequency of each trinucleotide mutation type in melanoma (**a**) and lung cancer (**e**). Unsupervised hierarchical clustering was used to define clusters based on the trinucleotide mutation signature in each sample. The promoter DHS/±1 kb flank mutation density ratio for each sample is shown at the leaf of the dendrogram, in which red and white depicts the highest and lowest ratio, respectively. **b**, **f**, Average mutation signature within each cluster displayed as bar graphs showing the percentage of total mutations attributed to each of the 96-trinucleotide mutation classes. **c**, **g**, Correlation of mutation signature against mutation signatures defined previously[6] with associated promoter mutation density ratio for each sample. **d**, **h**, Distribution of promoter mutation density ratio within each cluster. $***P < 0.001$, $**P < 0.01$ (unpaired *t*-test).

be attributed to the increased propensity for CPD formation at methylated CpGs[24,25], with CpG methylation generally being lower across TSSs. Indeed, the mutation profile closely mirrors that of the methylation profile over the TSS (Extended Data Fig. 6e), and the mutation density at [C/T]pG sites is significantly positively correlated with the methylation status of the respective promoter region (Extended Data Fig. 6f). Taken together, these results provide evidence that differential NER within the DHS and its flanking region underlies increased mutation density in the DHS of active gene promoters.

If differential NER underlies localized increased promoter mutation density, we reasoned that this observation would be most evident in cancers that rely on NER to repair DNA lesions. To test this, we clustered all melanoma samples based on their genome-wide trinucleotide mutation signature. Two distinct clusters were evident (Fig. 4a), corresponding to a non-ultraviolet (cluster 1) and an ultraviolet (cluster 2) damage signature (Fig. 4b, c). The promoter DHS/DHS flank mutation density ratio for samples with an ultraviolet signature was significantly higher than that in samples without an ultraviolet signature (Fig. 4d; $P < 0.001$, unpaired *t*-test), further implicating differential NER in generating promoter mutation hotspots. Promoter mutation density was also assessed in the context of trinucleotide mutation signatures in lung cancer, in which NER is also known to have a key role in the repair of smoking-induced DNA lesions, such as benzo[*a*]pyrene diol epoxide-DNA adducts[26]. Two distinct clusters were present (Fig. 4e) corresponding to samples with (cluster 1) and without (cluster 2) tobacco smoke damage signatures (Fig. 4f, g). Samples with smoking signatures had a significantly higher promoter DHS/DHS flank mutation density ratio compared with non-smoking signature samples (Fig. 4h; $P < 0.01$, unpaired *t*-test). These data suggest that impairment of NER is also responsible for promoter mutation hotspots in smoking-associated lung cancers (see Supplementary Data for discussion of promoter mutation hotspots in other cancer types).

Our results suggest an interaction exists between transcription initiation and NER, leading to highly localized mutation density, particularly at active gene promoters (Extended Data Fig. 7). From an evolutionary perspective, it is intriguing for a transcription-initiation-coupled mechanism to be present that permits accelerated accumulation of mutations at gene promoters. However, as many of the cancer genomes that we analysed contain tens to hundreds of mutations within promoter DHSs (Supplementary Table 2), promoter activity may in fact be reasonably robust to the effects of point mutations, with many genes possessing multiple promoters[27]. Nevertheless, as causal mutations can undoubtedly occur at gene promoters[28,29], our study highlights the need for careful scrutiny of the role of gene promoter mutations in cancer development.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 12 June 2015; accepted 19 February 2016.

1. Fredriksson, N. J., Ny, L., Nilsson, J. A. & Larsson, E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nature Genet.* **46,** 1258–1263 (2014).
2. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nature Genet.* **46,** 1160–1165 (2014).
3. Melton, C., Reuter, J. A., Spacek, D. V. & Snyder, M. Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nature Genet.* **47,** 710–716 (2015).
4. Poulos, R. C. *et al.* Systematic screening of promoter regions pinpoints functional cis-regulatory mutations in a cutaneous melanoma genome. *Mol. Cancer Res.* **13,** 1218–1226 (2015).
5. Hu, J., Adar, S., Selby, C. P., Lieb, J. D. & Sancar, A. Genome-wide analysis of human global and transcription-coupled excision repair of UV damage at single-nucleotide resolution. *Genes Dev.* **29,** 948–960 (2015).
6. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).
7. Mellon, I., Spivak, G. & Hanawalt, P. C. Selective removal of transcription-blocking DNA damage from the transcribed strand of the mammalian DHFR gene. *Cell* **51,** 241–249 (1987).
8. Zheng, C. L. *et al.* Transcription restores DNA repair to heterochromatin, determining regional mutation rates in cancer genomes. *Cell Rep.* **9,** 1228–1234 (2014).
9. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463,** 191–196 (2010).
10. Pleasance, E. D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463,** 184–190 (2010).
11. Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342,** 1235587 (2013).
12. Liu, L., De, S. & Michor, F. DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. *Nature Commun.* **4,** 1502 (2013).
13. Polak, P. *et al.* Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. *Nature Biotechnol.* **32,** 71–75 (2014).
14. Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488,** 504–507 (2012).
15. Woo, Y. H. & Li, W. H. DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. *Nature Commun.* **3,** 1004 (2012).
16. Supek, F. & Lehner, B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* **521,** 81–84 (2015).
17. Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518,** 360–364 (2015).
18. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507,** 455–461 (2014).
19. De Santa, F. *et al.* A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS Biol.* **8,** e1000384 (2010).
20. Koch, F. *et al.* Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nature Struct. Mol. Biol.* **18,** 956–963 (2011).
21. Tommasi, S., Oxyzoglou, A. B. & Pfeifer, G. P. Cell cycle-independent removal of UV-induced pyrimidine dimers from the promoter and the transcription initiation domain of the human *CDC2* gene. *Nucleic Acids Res.* **28,** 3991–3998 (2000).
22. Tu, Y., Tornaletti, S. & Pfeifer, G. P. DNA repair domains within a human gene: selective repair of sequences near the transcription initiation site. *EMBO J.* **15,** 675–683 (1996).
23. Tornaletti, S. & Pfeifer, G. P. UV light as a footprinting agent: modulation of UV-induced DNA damage by transcription factors bound at the promoters of three human genes. *J. Mol. Biol.* **249,** 714–728 (1995).
24. Rochette, P. J. *et al.* Influence of cytosine methylation on ultraviolet-induced cyclobutane pyrimidine dimer formation in genomic DNA. *Mutat. Res.* **665,** 7–13 (2009).
25. Cannistraro, V. J., Pondugula, S., Song, Q. & Taylor, J. S. Rapid deamination of cyclobutane pyrimidine dimer photoproducts at TCG sites in a translationally and rotationally positioned nucleosome *in vivo*. *J. Biol. Chem.* **290,** 26597–26609 (2015).
26. Gunz, D., Hess, M. T. & Naegeli, H. Recognition of DNA adducts by human nucleotide excision repair. Evidence for a thermodynamic probing mechanism. *J. Biol. Chem.* **271,** 25089–25098 (1996).
27. The FANTOM Consortium and the RIKEN PMI and CLST (DGT). A promoter-level mammalian expression atlas. *Nature* **507,** 462–470 (2014).
28. Horn, S. *et al. TERT* promoter mutations in familial and sporadic melanoma. *Science* **339,** 959–961 (2013).
29. Huang, F. W. *et al.* Highly recurrent *TERT* promoter mutations in human melanoma. *Science* **339,** 957–959 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements The authors thank TCGA, ICGC as well as numerous other groups who have made their data available for public analysis. The authors additionally thank members of Intersect Pty Ltd for providing high-performance computing resources and data storage used in this study. This work was funded by Cancer Institute NSW (13/DATA/1-02) and the Cure Cancer Foundation Australia with the assistance of Cancer Australia, through the Priority-driven Collaborative Cancer Research Scheme (APP1057921) to J.W.H.W. D.P. is supported by a UNSW Australia post-graduate scholarship, R.C.P is supported by an Australian Postgraduate Award, D.B. is supported by a National Health and Medical Research Council Early Career Fellowship (APP1073768), J.E.P. is funded by the National Health and Medical Research Council (Australia) and J.W.H.W. is supported by an Australian Research Council Future Fellowship (FT130100096).

Author Contributions Project planning and design: J.E.P. and J.W.H.W. Method design and data analysis: D.P., R.C.P., A.S., D.B. and J.W.H.W. Manuscript writing and figures: D.P., R.C.P. and J.W.H.W. All authors reviewed and edited the final manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.W.H.W. (jason.wong@unsw.edu.au).

METHODS

Data sets used and data processing. Somatic mutation data were obtained from four publicly available data sources: TCGA, ICGC, Alexandrov *et al.*[6] and Zheng *et al.*[8]. The single base substitution data from the ICGC were obtained from the ICGC data portal (release 16), and data from Alexandrov *et al.*[6] were obtained from ftp://ftp.sanger.ac.uk/pub/cancer/AlexandrovEtAl. For the XPC^{wt} and $XPC^{-/-}$ skin SCC samples, mutations were obtained from the database of Genotypes and Phenotypes (dbGap) (phs000830). These mutations were used directly for the analysis. For samples obtained from TCGA, mutations were called from BAM files obtained from CGHub[30] using Strelka[31] with default parameters. In all analyses, only single nucleotide variants (generally referred to as mutations in the manuscript) have been used, as the frequencies of other types of mutations were too low for robust statistical analyses.

Defining genomic regions and mutation densities. Enhancer and promoter DHSs were defined using cell-type matched DNase-seq data obtained from various publicly available data sources (Supplementary Table 1). Aligned DHS and histone ChIP-seq data were downloaded and peaks were called using the FindPeaks tool within the Homer package[32] using the 'dnase' and 'histone' modes, respectively. Putative promoter and enhancer DHSs were then called based on the overlap of DHS peaks (which are all 150 bp wide by default), with H3K4me3 and H3K4me1 peaks, respectively. To further increase the confidence of promoter and enhancer DHS annotations, only those that were identified by CAGE data as transcribed p1 promoters and enhancers respectively by FANTOM5 (refs 18, 27) were selected for the analysis. This resulted in a final set of promoter and enhancer DHS of 150 bp in size for each cell type, with each associated with a cancer type as identified in Supplementary Table 1. A universal promoter and enhancer DHS data set was also generated by merging all promoter and enhancer DHS regions across all cell types, but retaining only a single representative 150 bp DHS for regions where multiple DHSs overlapped. For DHS flanking regions, a 1 kb region on either side of the 150 bp DHS centre was used. Heterochromatin regions were regions that were identified as such by ChromHMM[33] across all cell types. For defining coding regions, intronic regions and TSSs, the canonical UCSC genes data set was obtained from the UCSC table browser. Overlapping coding regions were merged to generate a non-redundant set of coding regions for assessing coding mutation density. To exclude genomic regions where unique short read mapping may be challenging, *Duke Uniqueness*, *Duke Excluded Regions* and *DAC Blacklisted Regions* were obtained from the UCSC table browser and removed from all of the above annotations.

A summary of mutation counts for each sample and associated region sizes can be found in Supplementary Table 2. The mutation density of each region was reported as the number of mutations found in a particular genomic region, normalized by region size and the number of cancer samples. It was then converted to mutations per Mb.

A bootstrapping analysis was performed to determine the confidence interval threshold for the number of mutations required to identify a 2-fold increase in mutation density at promoters relative to flanking regions. For this analysis, mutations within each sample were randomly shuffled to other genomic locations and the density of mutation within promoter DHS and flanking regions was computed. This was repeated 1,000 times to obtain robust confidence intervals.

To assess statistically significant differences between promoter DHSs, enhancer DHSs and flanking regions, the χ^2 test with Yates correction was used. For evaluating the significance of increased local promoter DHS density across individual samples of a given tumour type, a paired ratio *t*-test was used. The results of these tests are summarized in Extended Data Tables 1 and 2. The investigators were not blinded to allocation during experiments and outcome assessment.

Trinucleotide mutation signature analysis. Trinucleotide mutation frequencies for each cancer sample were counted as previously described[6]. In brief, the frequency of point mutations was calculated in each of the possible 96 trinucleotide 5′ to 3′ contexts. These were counted either across the genome or within promoter DHSs or promoter DHSs ±1 kb flanking regions. Mutation signatures were defined by the relative frequency of these 96 trinucleotide mutations. Hierarchical clustering was performed using the mutation signatures to distinguish samples based on mutation processes[6]. Pearson's correlation was used to determine the association of mutation signatures from ref. 6 to the mutation signature of specific samples.

Normalization of sequence GC content and trinucleotide mutation signatures. The total numbers of C and G bases, or each of the 32 possible trinucleotides, were counted within each of the respective genomic regions. For normalization by GC content, the C > N and T > N mutation density was normalized by the percentage GC within each region. For normalization by trinucleotide mutation frequencies, each of the 96 possible trinucleotide mutations were normalized by the respective trinucleotide frequency within each region. The normalized mutation rate was then calculated from the sum of the normalized trinucleotide mutation frequencies for each region. The normalized ratio is thus the ratio of the sum of the normalized frequencies in the two regions.

Regression analysis. To standardise a set of promoters for regression analysis, we defined a set of core promoter regions as −100 bp of the TSS of canonical UCSC genes. This TSS −100 bp region was chosen as it reflects the most nucleosome-free region in gene promoters. For each of these regions, the following was computed, (1) DNase-seq coverage using the respective DNase-seq data set as defined in Supplementary Table 1; (2) gene expression of the associated gene based on the average expression of each gene across all samples of the corresponding cancer type with normalized RNA-seq expression data obtained from the ICGC data portal; (3) replication timing, calculated as the average value within the region from the wavelet-smoothed signal obtained from the ENCODE project[34]. The closest matching cell line for each cancer type was used (melanoma: NHEK, ovarian: HeLa-S3, lung: IMR90); (4) the proportion of rare SNPs, as the ratio of rare SNPs (defined as derived allele frequency <0.5%) to all SNPs from the 1000 Genomes Project[35]; (5) cancer genes, as only those listed by the Cancer Gene Census[36]; (6) the conservation of each promoter, as the average GERP score[37] across the region; (7) the number of mutations within each promoter for each cancer type analysed; (8) GC content of the region computed as described above; (9) relative frequency of each of the 32 trinucleotide combinations. Regression models were only computed for the mutations from melanoma, ovarian and lung cancers as only these cancers had sufficient numbers of total promoter mutations to allow for the generation of a reliable regression model and they also constituted the cancers in which a majority of samples exhibited increased local promoter DHS mutation density.

For the univariate logistic regression, with the exception of the cancer gene variable which was defined as a categorical variable, all variables were standardised to a mean of zero and standard deviation of one. The regression was performed using the *glm* package in *R*. The odds ratio was calculated by exponentiation of the coefficients and the *P* values were obtained directly from the regression model. For multivariate logistic regression, all variables were combined in a linear equation. The matrices used for regression analysis are provided in Supplementary Tables 3–5.

To assess statistical significance, we used Poisson regression to evaluate the associations between DNase I hypersensitivity or NER with mutation density. For DNase I hypersensitivity with NER, linear regression was used. The regression was performed using the *glm* package in *R*. The odds ratio was calculated by exponentiation of the coefficients and the *P* values were obtained directly from the regression model.

Generation of mutation and DHS profiles across transcription start sites. The set of TSSs were stratified into four quarters of DNase I hypersensitivity and expression using the data generated as described above for regression analysis. For each set of TSSs, mutation profiles were generated by counting the number of mutations for each respective cancer type across a ±5,000 bp window across each TSS. TSSs were orientated accordingly, such that the gene body is on the right of the TSS in the profiles. The mutation counts were normalized to mutations per Mb and plotted in 100 bp (for profiles stratified by DHS) and 5 bp (for profiles stratified by expression) windows.

Transcription factor footprinting analysis. Our logistic regression analysis revealed that the presence of promoter mutations was significantly anti-correlated with average conservation (GERP score)[37] of gene promoters in melanoma and lung cancer (Extended Data Table 3). Since transcription factor binding sites are generally more highly conserved than their flanking regions[38], we reasoned that the increased conservation of mutated bases may reflect an increased likelihood for somatic mutations to occur within transcription factor binding sites. To test this hypothesis transcription factor footprinting analysis was performed using melanocyte DGF data from the Human Epigenome Atlas (GSM1024610). Raw reads were obtained from the Sequence Read Archive (SRA) and reads aligned using BWA[39]. Since the transcription pre-initiation complex comprising of TATA-binding protein (TBP) and other general transcription factors has been shown to be present within a ~50-bp region upstream of TSSs of actively transcribed genes[40], Wellington[41] was used to compute default and 50 bp footprints (parameters: '-sh 13,17,1 -fp 46,54,1') across TSS −100 bp regions. The number of melanoma mutations was then counted inside and outside of footprints. To standardise the comparison between mutation counts in transcription factor binding (footprinted) and unoccupied (non-footprinted) sites, only promoter regions where a footprint had been detected were used for further analysis. For statistical analysis of significance, paired *t*-test was used for comparing mutated and non-mutated bases inside and outside of footprints. Mutation and DGF profiles were generated by averaging read coverage centred on all 50 bp footprints.

Matched DNase I hypersensitivity analysis. To directly compare the relative mutation density at promoter and enhancer DHSs, a set of active promoters

represented by those within the top 25% of DNase I hypersensitivity was selected. A corresponding set of enhancers with matched DNase I hypersensitivity was selected in which for each promoter, an enhancer within ±5 DNase-seq read coverage was randomly selected. The process was repeated 100 times to account for variations in enhancer selection.

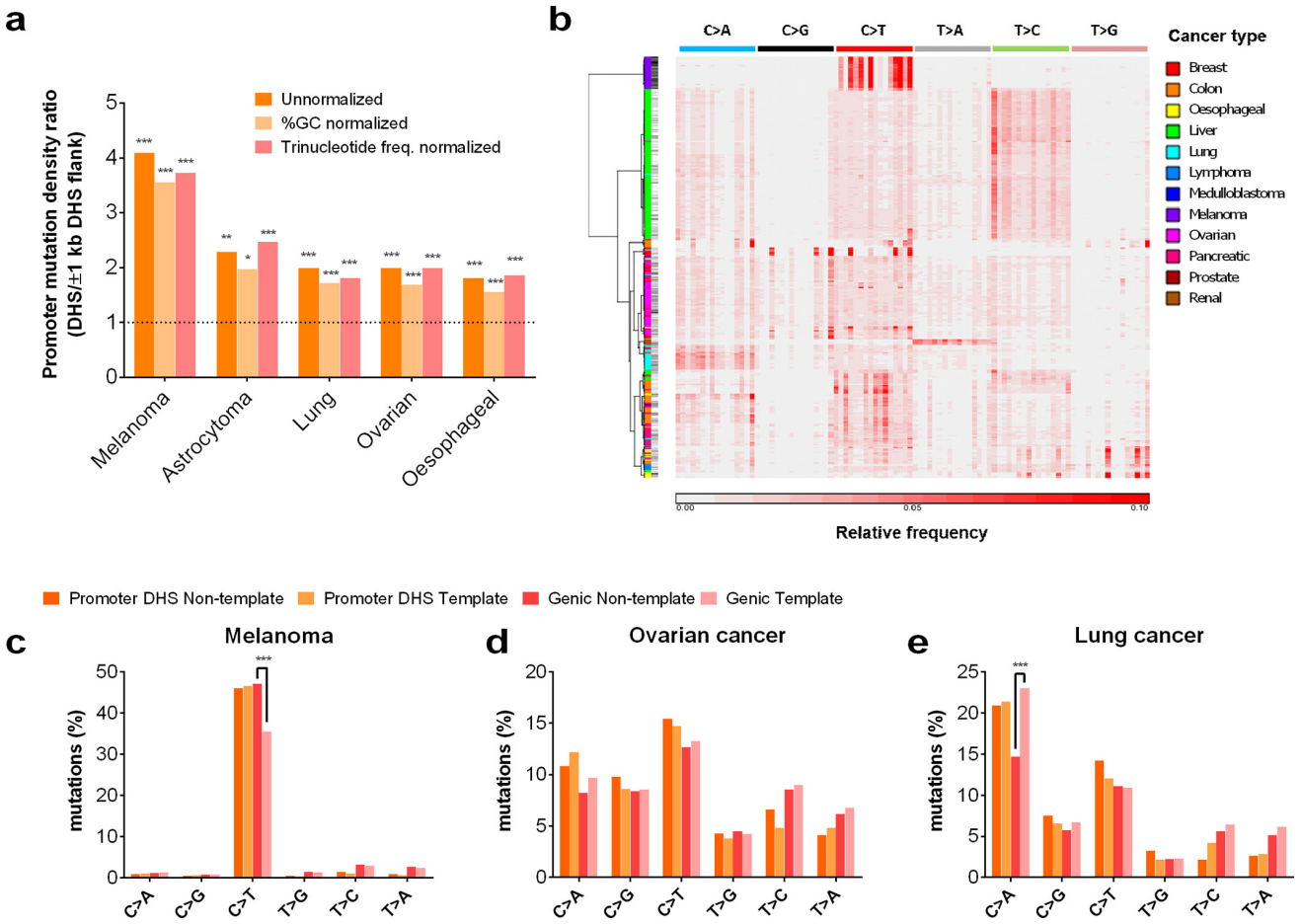
For the analysis of transcriptionally active versus less active enhancers, enhancer data from the FANTOM5 consortium was used[18]. FANTOM5 defined a set of ubiquitous enhancers that have been found to strongly promote the transcription of enhancer RNA across all cell lines examined[18]. The coordinates for ubiquitous enhancers ($n = 200$) and permissive enhancers ($n = 43{,}011$) were obtained from (http://enhancer.binf.ku.dk/presets/). To ensure that these sites do not overlap our promoter data set, we subtracted enhancer regions that overlapped with promoter DHS from any cell type. As for the comparison of active promoters and DNase I hypersensitivity matching enhancers described above, the same procedure was used to select DNase I hypersensitivity matching permissive enhancers. In this case, all ubiquitous enhancers were used as, by definition, they are active in all cell types.

NER profiles at gene promoters. Genome-wide NER sequencing (XR-seq) data sets of CPD and 6–4PP from ultraviolet-irradiated normal human skin fibroblast cells, generated previously[5], were obtained in SRA format (GSM1659156). The raw reads were trimmed using trimmomatic[42] and aligned using Bowtie[43] as described[5]. De-duplicated aligned reads for the two replicate experiments were then merged and average coverage profiles were generated. For the quantification of repair ratios against DNase I hypersensitivity at gene promoters, a 100 bp window was selected around all promoter DHS peak centres and flanking regions defined as ±150 bp. This range was selected as it was the region that best defined the peak and trough in NER near the TSS. As XR-seq reads are ~30 bp in length, to avoid multi-counting of reads that overlap both the DHS centre and flanking regions, the centre of each read was used to define its overlap with the genomic regions. The DNase I hypersensitivity of each promoter was quantified as the number of melanocyte DNase-seq reads overlapping the promoter DHS region. To establish the relationship between DNase I hypersensitivity and repair ratio, XR-seq reads for the respective regions were summed within bins of 25 DNase-seq coverage and normalized by region size. The repair ratio was calculated as the ratio of promoter DHS/flank for each bin. For enhancers, the repair ratio between the DHS centre and flank was similarly calculated for FANTOM5 ubiquitous and permissive enhancers.

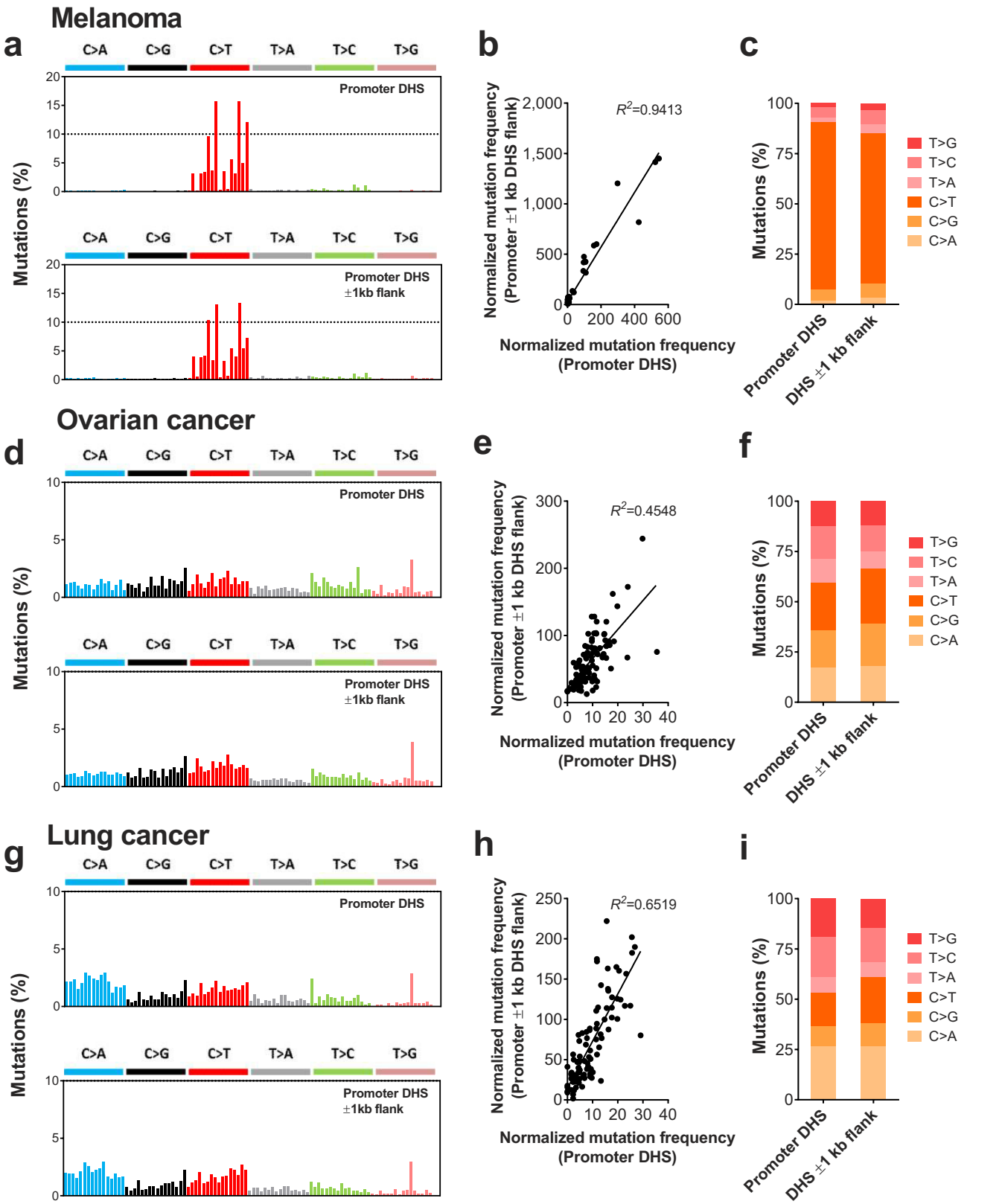
DNA methylation in normal human epithelial keratinocytes. To compare mutation density with CpG methylation status at gene promoters of $XPC^{-/-}$ SCC genomes, processed whole genome bisulfite sequencing data from Normal human epithelial keratinocytes (NHEK) was obtained from the Human Epigenome Atlas[44]. For the generation of the average methylation profile, deepTools[45] was used. To correlate promoter methylation status with the mutation density of C > T mutations within a [C/T]CpG context, the mutation density and average fraction of methylation was measured within all TSS ±1 kb regions.

Code availability. Scripts and annotation files used for data analysis are available as a zipped file under Supplementary Information.

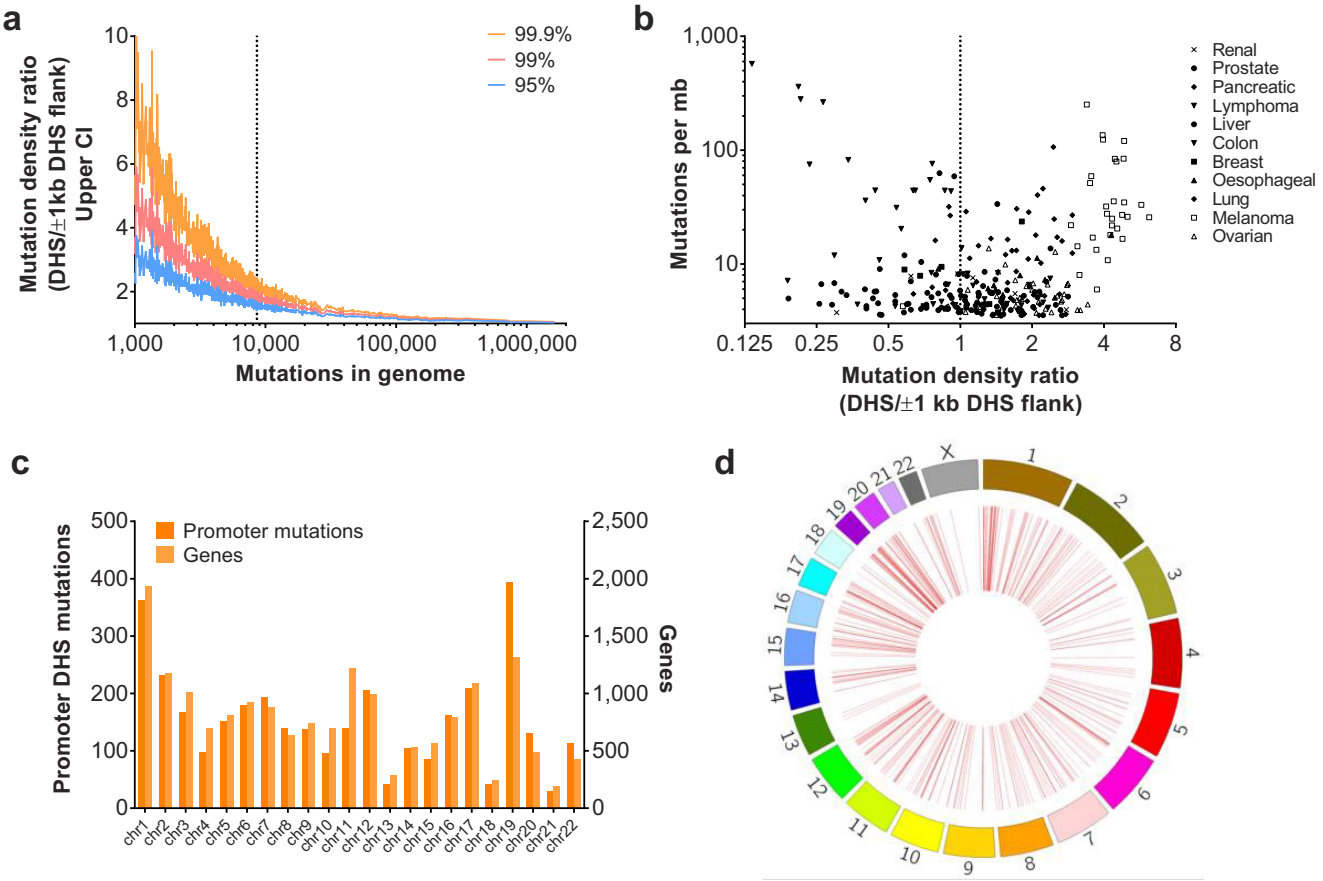
30. Wilks, C. *et al.* The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data. *Database* **2014**, bau093 (2014).
31. Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28,** 1811–1817 (2012).
32. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38,** 576–589 (2010).
33. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods* **9,** 215–216 (2012).
34. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
35. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
36. Futreal, P. A. *et al.* A census of human cancer genes. *Nature Rev. Cancer* **4,** 177–183 (2004).
37. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15,** 901–913 (2005).
38. Berezikov, E., Guryev, V., Plasterk, R. H. & Cuppen, E. CONREAL: conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting. *Genome Res.* **14,** 170–178 (2004).
39. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
40. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489,** 83–90 (2012).
41. Piper, J. *et al.* Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Res.* **41,** e201 (2013).
42. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).
43. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
44. Zhou, X. *et al.* The Human Epigenome Browser at Washington University. *Nature Methods* **8,** 989–990 (2011).
45. Ramírez, F., Dundar, F., Diehl, S., Gruning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42,** W187–W191 (2014).



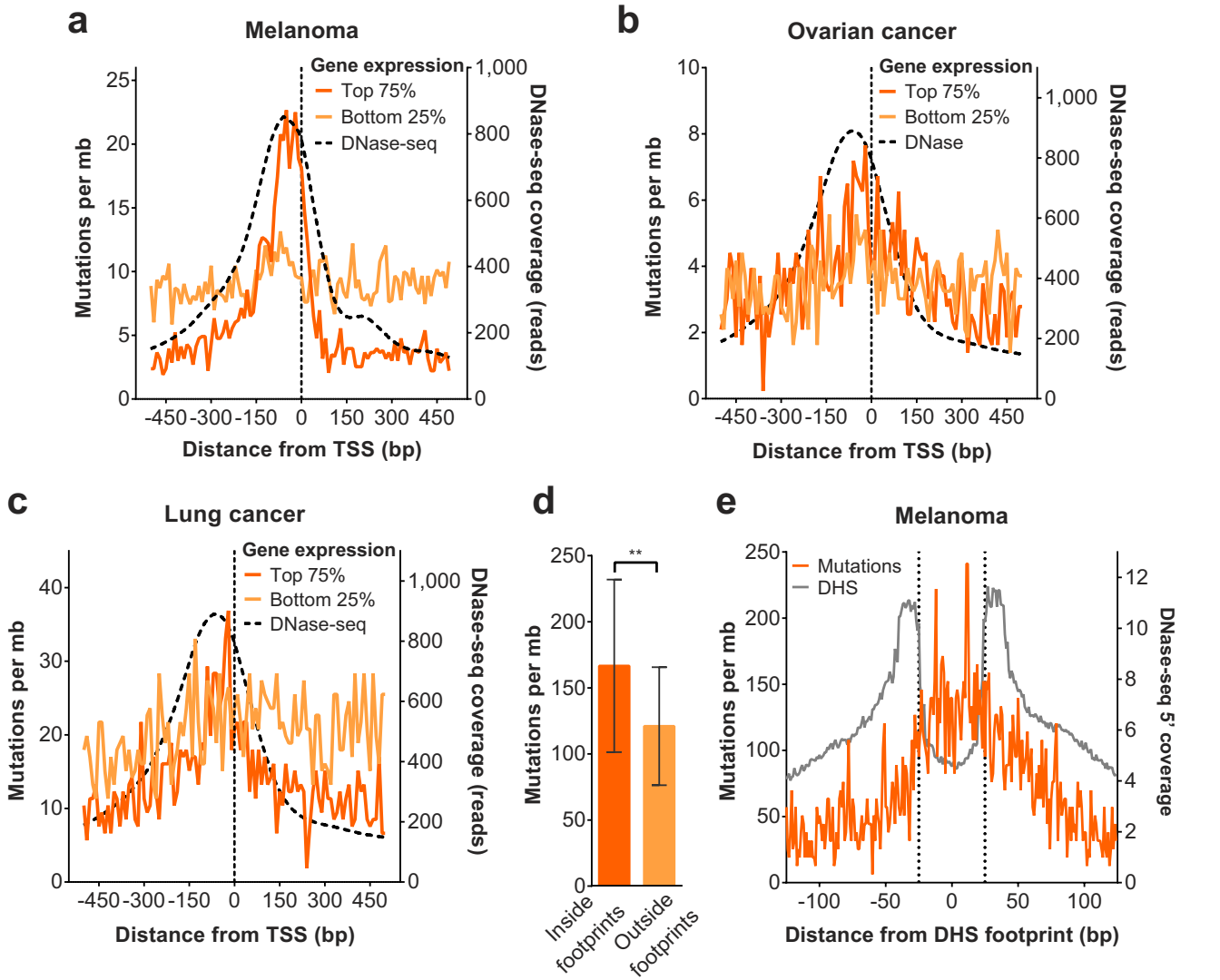
Extended Data Figure 1 | Relationship between sequence composition and trinucleotide mutation signatures in promoter DHS mutations. **a**, The mutation density ratio of promoter DHS/DHS flanking regions (DHS ±1 kb) in melanoma, astrocytoma, lung, ovarian and oesophageal cancers before and after adjustment by percentage GC content or trinucleotide frequencies within the respective regions. Adjustment was performed by dividing the mutation density in the promoter DHS and DHS flanking region by the percentage GC ratio or trinucleotide frequencies in the two regions, respectively. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$ (χ^2 test). **b**, Heatmap showing the relative frequency of each trinucleotide mutation signature across all samples with greater than 8,602 mutations (see Extended Data Fig. 3a). Unsupervised hierarchical clustering was used to define clusters based on the trinucleotide mutation signature of each sample. The promoter DHS/±1 kb flank mutation density ratio for each sample is shown at the leaf of the dendrogram where black and white depicts the highest and lowest ratios, respectively. The cancer type is colour-coded as defined by the key on the right of the figure. **c**–**e**, Mutations were separated into 6 classes and relative frequency was evaluated over promoter DHSs and genic regions in melanoma, ovarian and lung cancer according to the template and non-template strands relative to the associated gene. $^{***}P < 0.001$ (χ^2 test).



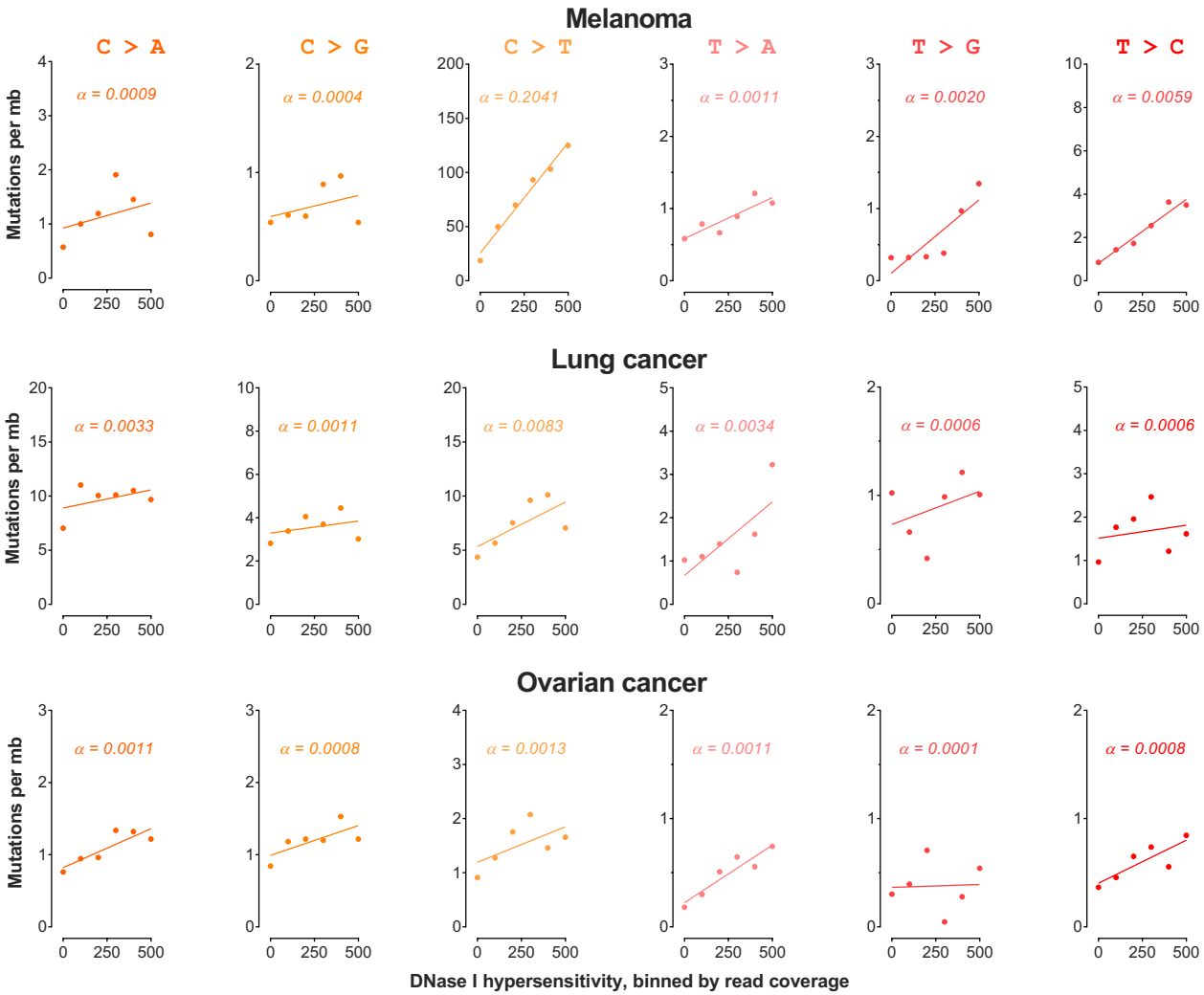
Extended Data Figure 2 | Comparison of mutation signatures in promoter DHS and its ±1-kb flanking region. a, **d**, **g**, Trinucleotide mutation signatures within promoter DHSs and ±1 kb flanking regions in melanoma, ovarian and lung cancer, respectively. All signatures have been normalized by trinucleotide frequencies within their respective regions. **b**, **e**, **h**, Correlation of the normalized trinucleotide mutation signature frequencies in the promoter DHS versus the ±1 kb flanking region in melanoma, ovarian and lung cancers, respectively. The Pearson's correlation was calculated by linear regression. **c**, **f**, **i**, Comparison of the distribution of each of the 6 mutation classes in promoter DHSs and ±1 kb flanking regions with mutation counts normalized by GC frequency. There are significantly more C > T mutations in melanoma ($P < 0.001$, χ^2 test) and more T > N mutations in ovarian ($P < 0.001$, χ^2 test) and lung cancers ($P < 0.001$, χ^2 test) based on mutation counts normalized by GC frequency.



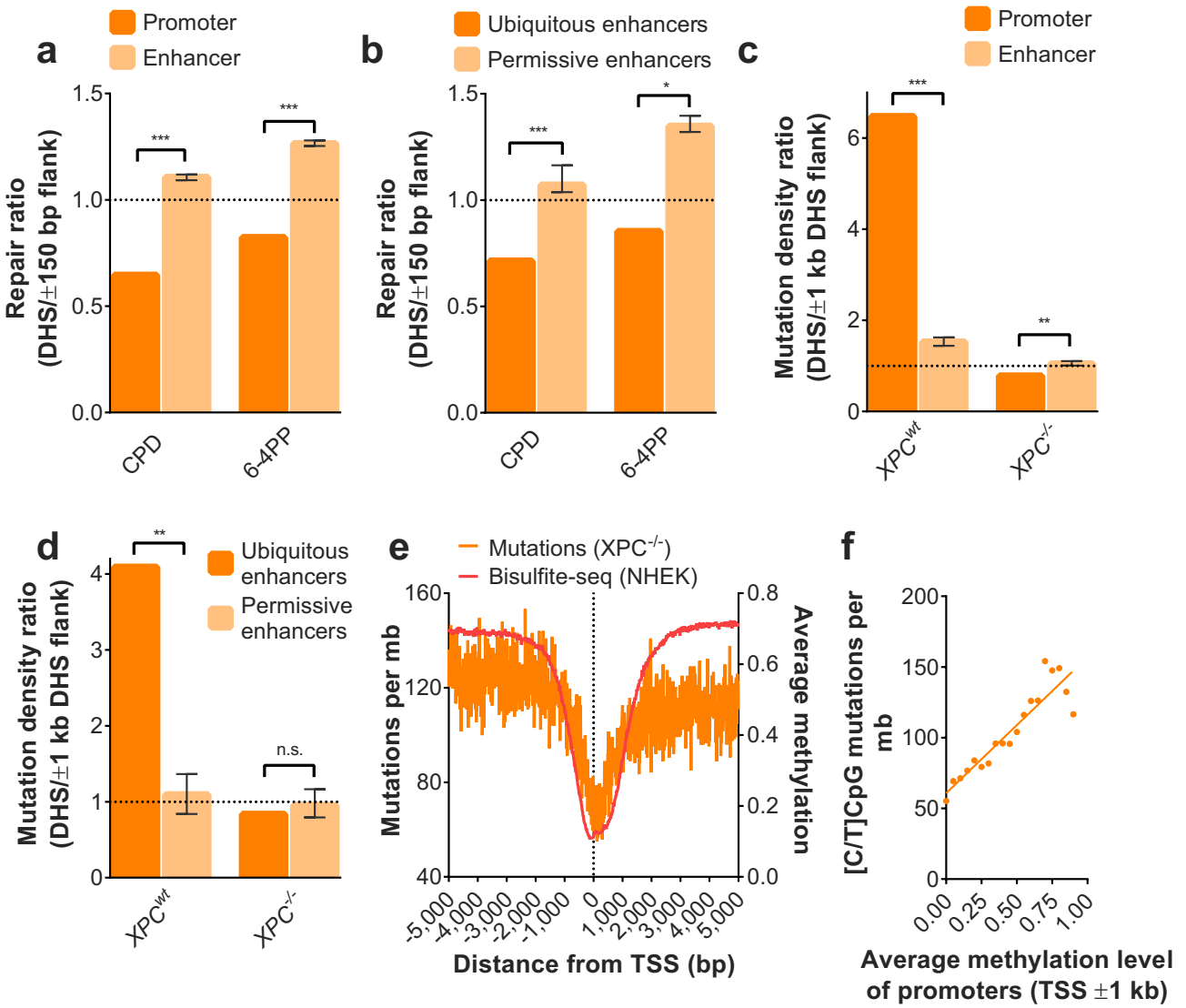
Extended Data Figure 3 | Distribution of promoter DHS mutations in relation to genome mutation load and across chromosomes. a, Mutation power analysis for detection of significant promoter DHS mutation enrichment. Bootstrapping analysis was performed to assess the number of mutations required in an individual sample to achieve >95% confidence (that is, less than 5% of resampling resulting in a >2-fold enrichment of promoter DHS mutation density relative to its (±1 kb) flanking region. The dotted line marks the number of mutations (8,602) required to detect >2-fold enrichment of promoter DHS mutations relative to flanking region with at least 99% confidence and this threshold was used to select cancer samples for individual analysis. **b**, The mutation density ratio of the promoter DHS/±1 kb DHS flanking region for individual cancer genomes with at least 8,602 mutations plotted against genomic mutation density. **c**, The number of promoter DHS melanoma mutations and the number of genes within each chromosome. **d**, Circos plot showing the location of promoter DHS mutations (red lines) in TCGA melanoma sample TCGA-EE-A3J5.



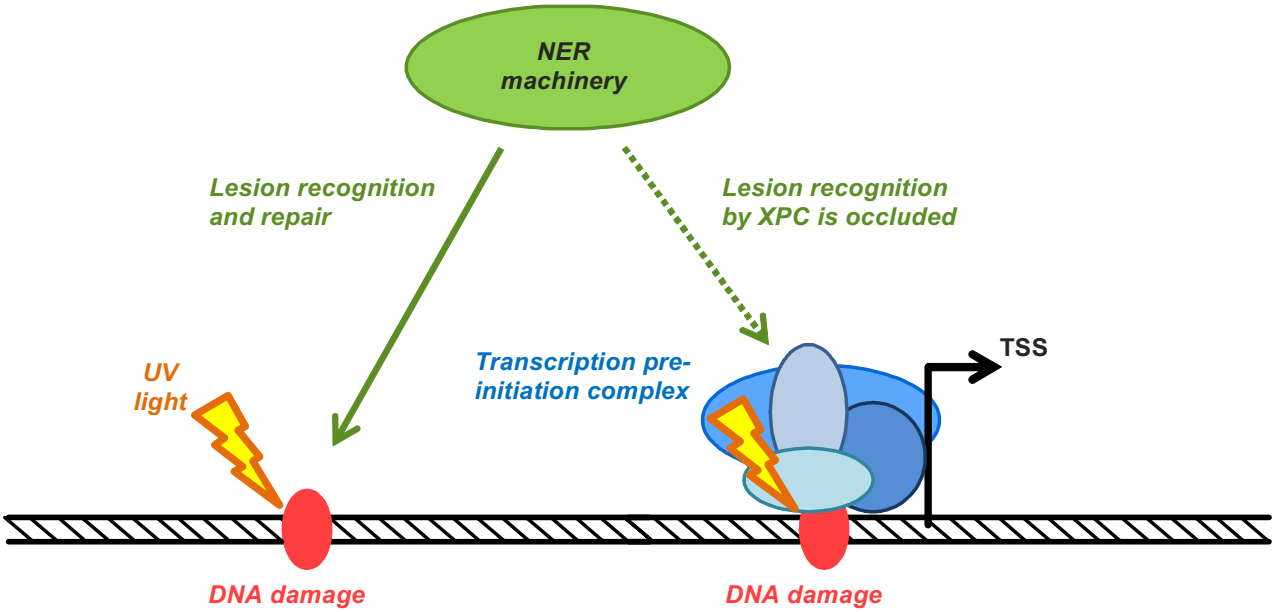
Extended Data Figure 4 | Mutation density is increased across the DHS centre of highly expressed genes. a–**c**, Melanoma (**a**), ovarian (**b**) and lung (**c**) cancer mutation profiles and melanocyte (**a**), ovary cell (**b**) and A549 cell (**c**) DNase-seq cleavage profiles, respectively, centred around the TSS. Profiles for mutations were stratified by quartiles of gene expression while DNase-seq profiles were averaged across all genes. Mutation profiles were smoothed using 5 bp windows. Mutation density profiles are oriented according to strand. **d**, Mutation density in melanoma within and outside of digital genomic footprints within TSS −100 bp promoter regions of melanocytes. Mean mutation density is shown, together with 95% confidence intervals across all 36 samples. Footprinted regions represent transcription factor (TF) bound sites whereas non-footprinted regions represent unoccupied sites (**$P < 0.01$, paired t-test). **e**, Melanoma mutation and melanocyte DNase-seq cleavage profiles centred around 50 bp footprints identified within TSS −100 bp promoter regions. Mutation profiles were smoothed using 5 bp windows.



Extended Data Figure 5 | Association between mutation density of the 6 mutation classes against chromatin accessibility. Chromatin accessibility was measured by DNase-seq read coverage in bins of 100 bp. Slope (α) was calculated from the linear regression of the binned data.



Extended Data Figure 6 | Comparison of promoter with enhancer mutation density and relationship between mutation density and DNA methylation in *XPC*$^{-/-}$ skin cancer. a, **b**, CPD and 6–4PP XR-seq repair density ratio of DNase I hypersensitivity matched (**a**) active promoters and enhancers, and (**b**) ubiquitous and permissive enhancers relative to their DHS flanking regions. For promoters, a set of active promoters represented by the top 25% of nucleosome free promoters based on melanocyte DNase-seq data was used. A corresponding set of enhancers of equal size were selected with matching DNase-seq coverage. The error bar on the enhancer data set shows the interquartile range of repair ratios over 100 randomized samplings of enhancers with matching DNase-seq coverage. For the comparison of ubiquitous and permissive enhancers, the full set of ubiquitous enhancers from FANTOM5 ($n = 200$) were used and a matching set of equally nucleosome free permissive enhancers were sampled and repeated 100 times as described above (also see Methods). SCC *XPC*^{wt} and *XPC*$^{-/-}$ mutation density ratio of DNase I hypersensitivity matched (**c**) active promoters and enhancers, and (**d**) ubiquitous and permissive enhancers relative to their DHS flanking regions. Promoter, enhancer, ubiquitous enhancer and permissive enhancer regions were generated as described for XR-seq data in **a** and **b**. **e**, Mutation density and methylation profile ±5 kb of the TSS of genes for *XPC*$^{-/-}$ SCC and normal human epithelial keratinocytes (NHEK), respectively. Methylation profiles were generated using the fraction methylation data calculated using whole genome bisulfite sequencing (bisulfite-seq) data from the Human Epigenome Atlas. **f**, Association between the average methylation level of gene promoters (TSS ±1 kb) and the density of [C/T]CpG mutations in *XPC*$^{-/-}$ SCC. The average methylation level of each promoter region was calculated using the mean fraction methylation of each [C/T]CpG within the region as measured by bisulfite-seq in NHEK cells. For **a**–**d**: $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, n.s., not significant (χ^2 test).



Extended Data Figure 7 | Schematic diagram of proposed mechanism leading to localized increased promoter mutation density. DNA damage (such as CPD or 6–4PP) caused by ultraviolet (UV) irradiation is typically recognized by NER machinery such as *XPC*, to initiate DNA repair (left). In highly transcribed promoters, the transcription pre-initiation complex prevents repair machinery such as *XPC* from recognizing the DNA lesion (right), leaving it unrepaired and ultimately leading to mutation formation upon DNA replication.

Extended Data Table 1 | Comparison of promoter and enhancer mutation densities across 14 cancer types

cancer type	# of samples	# of promoter mutations	# of enhancer mutations	promoter region size	enhancer region size	promoter mutation density	enhancer mutation density	promoter/ enhancer density ratio	X^2 p-value*
Melanoma	36	4,172	784	1,859,126	1,334,525	62.3351	16.3187	3.8198	<0.0001
Lung	24	1,038	371	1,847,950	1,231,650	23.4043	12.5509	1.8647	<0.0001
CLL	28	34	16	2,036,936	1,707,159	0.5961	0.3347	1.7810	0.0736
Astrocytoma	100	17	9	1,879,722	1,757,700	0.0887	0.0502	1.7663	0.2285
Ovarian	93	935	322	1,846,213	985,950	5.4456	3.5117	1.5507	<0.0001
Lymphoma	44	183	80	1,694,595	1,094,550	2.4543	1.6611	1.4775	0.0041
Oesophageal	16	115	38	1,898,301	850,950	3.7863	2.7910	1.3566	0.1213
Pancreatic	201	686	231	1,550,010	619,800	2.2019	1.8542	1.1875	0.0261
Breast	119	448	385	1,732,115	1,636,084	2.1735	1.9775	1.0991	0.1846
Renal	95	198	76	1,818,474	595,350	1.1461	1.3437	0.8529	0.2674
Liver	244	1,041	459	1,676,855	595,350	2.5443	3.1597	0.8052	0.0001
Medulloblastoma	100	63	19	1,828,761	426,600	0.3445	0.4454	0.7735	0.3994
Prostate	14	28	16	1,733,901	552,845	1.1535	2.0672	0.5580	0.0870
Colon	47	652	832	1,738,451	850,950	7.9797	20.8028	0.3836	<0.0001

*χ^2 test with Yates correction for the difference between promoter and enhancer mutation density for each cancer type.

Extended Data Table 2 | Comparison of promoter DHS and promoter DHS flank (±1 kb) mutation density across 14 cancer types

cancer type	# of samples	# of promoter mutations	# of promoter flank (+/-1 kb) mutations	promoter region size	promoter flank region size	promoter mutation density	promoter flank mutation density	promoter/flank density ratio	χ^2 test p-value*	paired ratio *t*-test p-value†
Melanoma	36	4,172	13,590	1,859,126	24,819,900	62.3351	15.2096	4.0984	<0.0001	<0.0001
Astrocytoma	100	17	99	1,879,722	25,096,101	0.0887	0.0387	2.2926	0.0021	NA
Lung	24	1,038	6,937	1,847,950	24,655,919	23.4043	11.7230	1.9964	<0.0001	<0.0001
Ovarian	93	935	6,268	1,846,213	24,625,652	5.4456	2.7369	1.9897	<0.0001	<0.0001
Oesophageal	16	115	848	1,898,301	25,340,056	3.7863	2.0916	1.8103	<0.0001	0.0120
Prostate	14	28	235	1,733,901	23,144,202	1.1535	0.7253	1.5904	0.0264	NA
Pancreatic	201	686	6,655	1,550,010	20,694,627	2.2019	1.5999	1.3763	<0.0001	<0.0001
Breast	119	448	4,760	1,732,115	23,124,510	2.1735	1.7298	1.2565	<0.0001	NA
Medulloblastoma	100	63	707	1,828,761	24,423,659	0.3445	0.2895	1.1901	0.2102	NA
Liver	244	1,041	12,635	1,676,855	22,371,574	2.5443	2.3147	1.0992	0.0035	0.5153
Renal	95	198	2,509	1,818,474	24,274,013	1.1461	1.0880	1.0534	0.5234	0.6906
Lymphoma	44	183	2,547	1,694,595	22,623,509	2.4543	2.5587	0.9592	0.6101	0.5697
CLL	28	34	494	2,036,936	27,177,647	0.5961	0.6492	0.9183	0.6892	NA
Colon	47	652	21,316	1,738,451	23,203,714	7.9797	19.5457	0.4083	<0.0001	0.0057

*χ^2 test with Yates correction for the combined mutations for each cancer type.
†Paired ratio *t*-test was used to determine significance of samples within cancer types.
NA indicates cancers in which there were fewer than 2 individual samples with a genome mutation load of more than 8,602, and as a result were not tested by paired ratio *t*-test.

Extended Data Table 3 | Logistic regression analysis of mutated promoters (TSS −100 bp) in melanoma, ovarian and lung cancer against various genetic and epigenetic characteristics known to contribute to variations in mutation density in the genome

Melanoma

	OR	95% CI	p-value	adj. OR	adj. 95% CI	adj. p-value
DNase I coverage	1.51	1.45,1.56	< 2E-16	1.51	1.45,1.56	<2E-16
Gene expression	1.08	1.04,1.13	0.00032	-	-,-	-
GC content	1.04	1.00,1.08	0.0836	1.04	0.99,1.09	0.1536
Replication timing	1.04	0.99,1.08	0.0937	0.95	0.91,0.99	0.0195
Proportion rare SNP	0.97	0.95,1.00	0.0849	0.99	0.96,1.01	0.2854
Cancer gene	0.9	0.69,1.16	0.436	0.82	0.63,1.07	0.1501
Conservation (GERP)	0.85	0.82,0.89	1.08E-13	0.88	0.84,0.92	3.10E-09

Lung Cancer

	OR	95% CI	p-value	adj. OR	adj. 95% CI	adj. p-value
DNase I coverage	1.09	1.03,1.16	0.00321	1.12	1.06,1.19	0.000171
Gene expression	1.03	0.97,1.10	0.376	-	-,-	-
GC content	1.18	1.11,1.27	1.18E-06	1.24	1.15,1.34	1.86E-08
Replication timing	0.87	0.82,0.93	2.10E-05	0.82	0.76,0.87	7.76E-10
Proportion rare SNP	1	0.97,1.04	0.66	0.99	0.93,1.03	0.5912
Cancer gene	1.21	0.82,1.72	0.321	1.19	0.80,1.70	0.365965
Conservation (GERP)	0.91	0.85,0.98	0.00867	0.92	0.86,0.98	0.015019

Ovarian Cancer

	OR	95% CI	p-value	adj. OR	adj. 95% CI	adj. p-value
DNase I coverage	1.25	1.17,1.32	4.86E-13	1.25	1.17,1.33	3.08E-12
Gene expression	1.38	1.14,1.67	0.00101	-	-,-	-
GC content	1.1	1.02,1.18	0.0149	1.13	1.04,1.23	0.00318
Replication timing	0.99	0.92,1.06	0.696	0.91	0.85,0.98	0.01465
Proportion rare SNP	1	0.96,1.03	0.988	0.99	0.95,1.02	0.65258
Cancer gene	1.34	0.89,1.94	0.142	1.22	0.81,1.78	0.31895
Conservation (GERP)	0.95	0.88,1.02	0.131	0.96	0.95,1.02	0.23806

adj., adjusted values from multivariate regression model; OR, odds ratio. For the multivariate analysis, the trinucleotide frequencies were also included in the logistic regression model, the full model is shown in Supplementary Tables 6–8. For multivariate models, gene expression has been excluded as its values are highly correlated with DNase I coverage.

LETTER

doi:10.1038/nature17661

Nucleotide excision repair is impaired by binding of transcription factors to DNA

Radhakrishnan Sabarinathan[1], Loris Mularoni[1], Jordi Deu-Pons[1], Abel Gonzalez-Perez[1] & Núria López-Bigas[1,2]

Somatic mutations are the driving force of cancer genome evolution[1]. The rate of somatic mutations appears to be greatly variable across the genome due to variations in chromatin organization, DNA accessibility and replication timing[2–5]. However, other variables that may influence the mutation rate locally are unknown, such as a role for DNA-binding proteins, for example. Here we demonstrate that the rate of somatic mutations in melanomas is highly increased at active transcription factor binding sites and nucleosome embedded DNA, compared to their flanking regions. Using recently available excision-repair sequencing (XR-seq) data[6], we show that the higher mutation rate at these sites is caused by a decrease of the levels of nucleotide excision repair (NER) activity. Our work demonstrates that DNA-bound proteins interfere with the NER machinery, which results in an increased rate of DNA mutations at the protein binding sites. This finding has important implications for our understanding of mutational and DNA repair processes and in the identification of cancer driver mutations.

The accumulation of somatic mutations in cells results from the interplay of mutagenic processes, both internal and exogenous, and mechanisms of DNA repair. Detailed early biochemical studies[7,8] and recent efforts to sequence the genomes of tumours from different cancer types[9,10] have shed light on this. Mutational signatures associated with various tumorigenic mechanisms have been identified across cancer types[11], and genomic features such as chromatin organization, DNA accessibility, and DNA replication timing[2–5] have been associated with the variation of somatic mutation rates at the megabase scale. Two recent studies proposed a causal relationship between the accessibility of chromosomal areas to the DNA repair machinery and their mutational burden. Supek and Lehner[12] pointed to variable repair of DNA mismatches as the basis of the megabase scale variation of somatic mutation rates across the human genome. Polak *et al.*[4] attributed lower somatic mutation rates at DNase I hypersensitive sites (DHS) than at their flanking regions and the rest of the genome in cell lines and primary tumours to higher accessibility to the global genome repair machinery. Similarly, nucleosome occupancy has been linked to regional mutation rate variation between nucleosome-bound DNA and linker regions[13–16], while two recent studies found a relation between transcription factor binding sites (TFBS) and nucleotide substitution rates. Reijns *et al.*[17] detected increased levels of nucleotide substitutions around TFBS in the yeast genome, which was attributed to DNA-binding proteins acting as partial barriers to the polymerase δ mediated displacement of polymerase α synthesized DNA. Katainen *et al.*[18] found that CTCF/cohesin-binding sites are frequently mutated in colorectal tumours and in a small subset of tumours of other cancer types, and suggested that these mutations are probably caused by challenged DNA replication under aberrant conditions.

To determine the impact of DNA-binding proteins on DNA repair, we analysed the somatic mutation burden at TFBS in the genomes of 38 primary melanomas sequenced by The Cancer Genome Atlas[10,19]. We found that the mutation rate was approximately five times higher in active TFBS, that is, those overlapping DHS (Fig. 1a) than in their flanking regions ($P < 2.2 \times 10^{-16}$, chi-square test). We determined that this elevated mutation rate could not be explained by the sequence context (Fig. 1a), and that it did not occur at inactive TFBS (Fig. 1a and Extended Data Fig. 1), indicating that it is directly related to the protein being bound to DNA. Furthermore, this enrichment for mutations appeared at the active binding sites of most transcription factors (TFs) (Fig. 1b, Extended Data Fig. 2 and Supplementary Table 1); the signal was discernible in most individual melanomas (Fig. 1c and Supplementary Table 2), and it increased with genome-wide mutation rate. Moreover, the signal was also apparent across the genome of a sample taken from normal human skin[20] (Fig. 1c), which indicates that the accumulation of mutations in TFBS results of a normal process rather than a pathogenic effect in tumour cells.

Most somatic mutations in melanocytes are caused by exposure to ultraviolet (UV) radiation[11]. UV radiation causes specific DNA lesions or DNA photoproducts of cyclobutane pyrimidine dimers (CPDs) and pyrimidine–pyrimidone (6–4) photoproducts (6–4PPs), at the sites of dipyrimidines[21]. As expected, C > T (G>A) mutations predominated in melanomas over other nucleotide changes (Fig. 1d), both within TFBS and at their flanks. This could be explained by either a faulty DNA repair[7,8] or a higher probability of UV induced lesions[22,23] at protein-bound DNA.

Next, we focused on active TFBS in distal regions from transcription start sites (TSS), and again found increased mutation rate at binding sites, flanked by periodic peaks of mutation rate observed at a distance of ∼170 bp, which coincides well with the size of the DNA wrapped around nucleosomes (∼146 bp) and the linker DNA, and could not be explained by sequence context (Fig. 2a). When we superimposed the nucleosomes positioning signals from the ENCODE data[24] and these mutation rate peaks, we verified that their positions matched well. Furthermore, we found that the peak of mutation rate observed at the centre of DHS regions occurred exclusively at TFBS and was absent from DHS sites without TFBS (Fig. 2b and Extended Data Fig. 3). This corroborated the model that whatever process was causing the increase in the mutation rate, it required that the proteins be bound to the DNA.

We then determined if the cause of the higher mutation rate in TFBS and nucleosomes was the reduced accessibility to the protein-bound DNA of the NER machinery. Non-repaired lesions would be bypassed by polymerases carrying out translesional DNA synthesis, thus resulting in mutations[25]. To test this, we assembled nucleotide-resolution maps of the NER activity of the two products of UV-induced DNA damage, CPDs and 6–4PPs, generated by ref. 6 using XR-seq in irradiated skin fibroblasts[6]. In XR-seq, the excised ∼30-mer around the site of damage generated during nucleotide excision repair is isolated and subjected to high-throughput sequencing. When we analysed the genome-wide signal of this NER map, we found a strong decrease in the amount of CPD and 6–4PP repair at the centre of TFBS (Fig. 3a and Extended Data Fig. 4a), compared to their flanking regions. The decrease was apparent both in wild-type cells (NHF1), and CS-B

[1]Research Program on Biomedical Informatics, IMIM Hospital del Mar Medical Research Institute and Universitat Pompeu Fabra, Doctor Aiguader 88, 08003 Barcelona, Spain. [2]Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain.

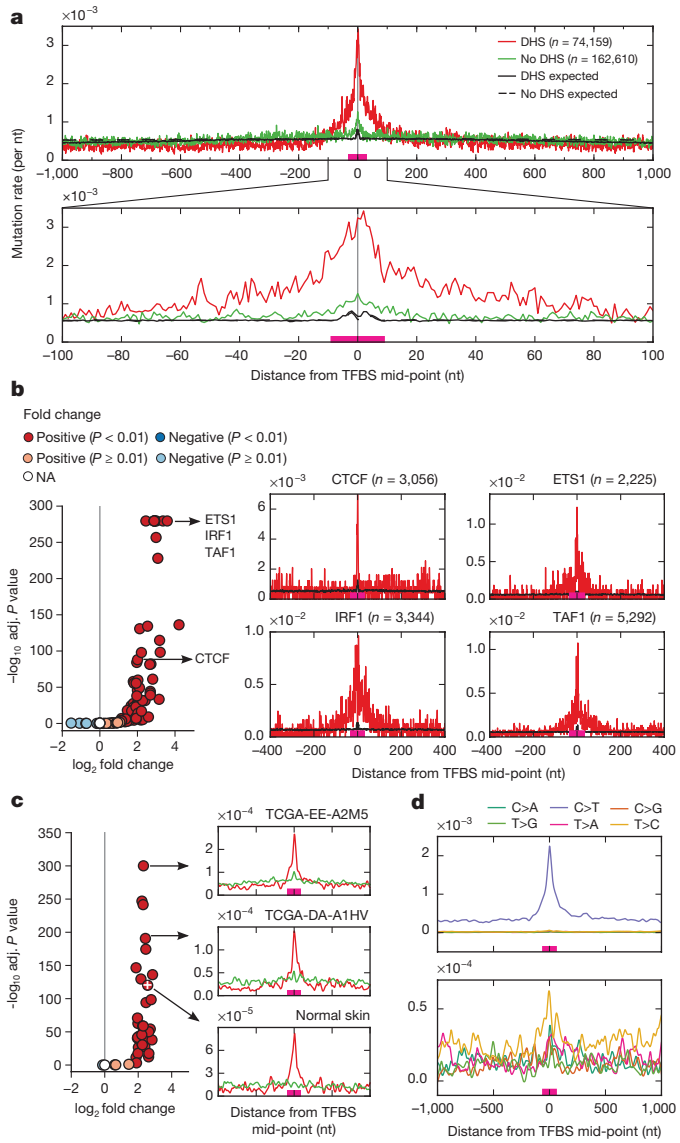


Figure 1 | Elevated mutation rate at TFBS in melanomas. a, Mutation rates are approximately fivefold higher within active TFBS, those overlapping DHS in melanocytes, than in flanking regions (red line). In contrast, non-active TFBS, those non-overlapping DHS in melanocytes, do not show increased mutation rates (green line). The high increase in mutation rate is not explained by sequence context; black lines show the expected mutation rate per position when distributing all observed mutations in the region according to the probability of mutations in different trinucleotide contexts. **b**, **c**, A significant increase in mutation rate in TFBS compared to flanking regions is observed for most individual transcription factors (**b**) and in most of the individual melanoma samples and a normal human skin sample (**c**). The $\log_2$ fold change (FC) on the *x* axis represents if the mutation rate in TFBS is higher (positive fold change) or lower (negative fold change) than the expected rate, and the corresponding significance value (from a chi-square test) is shown on the *y* axis for each transcription factor (**b**) or melanoma sample (**c**). **d**, The contribution of C > T mutations to mutational density is higher compared to the other mutation types. The zero coordinate on the *x* axis corresponds to the TFBS mid-point, and the magenta line above it represents the average size of TFBS.

mutant cell lines, which lack transcription-coupled repair[6] (Fig. 3a and Extended Data Fig. 4a), and it appeared at the binding sites of individual transcription factors (Extended Data Fig. 4b). Moreover, we found that the level of DNA excision repair (and the mutation rate) at TFBS correlated with the strength of their binding (Fig. 3b and Extended Data Fig. 5). We concluded from these observations that the higher mutation rate observed at active TFBS is caused by a decrease of the NER activity.

A previous study related higher DNA repair activity at DHS compared to outside these regions to greater accessibility to the repair machinery[4]. By specifically deconvoluting the signal of mutation rate within DHS, our work shows that bound TFs at the centre of DHS actually hinder DNA repair. This interplay of greater NER at DHS and lower NER at TF bound sites at their centre results in a 'volcano-shaped' pattern of NER activity around the TFBS, with a strong depletion exactly at its centre flanked by two mountains in the DHS area around it (Fig. 3). The volcano shape is more pronounced at distal TFBS, those that occur distant from transcription start sites (Fig. 3a), which may be explained by the presence of shorter regions of open chromatin surrounded by compacted DNA. Moreover, a periodicity in NER activity is observable for the first nucleosomes around TFBS (Fig. 3a), which matches well the previously noted periodical variation of the mutation rate. Also consistent with the mutation rate pattern, the signal of decreased NER activity is clearer at the centre of DHS-promoters-TFBS, exactly at the position of the TFBS (Extended Data Fig. 6). These results demonstrate that repair activity in DHS regions is in general higher than in non-DHS regions, supporting previous observations[4]; however, this activity is specifically impaired at sites with bound TFs.

NER consists of two pathways: global repair — targeting the lesions in a genome-wide manner — and transcription-coupled repair that recognizes lesions within transcribed regions[21]. These pathways differ in the initial steps of damage recognition, although they share the core component that excise damaged regions. To discern the effect of DNA bound TFs on transcription coupled NER we focused on transcribed regions centred at TFBS at least 200 bp downstream of TSS, and plotted together mutation rate and XR-seq data in XP-C cells, which only have transcription-coupled repair[6]. Mutation rate is also increased at the centre of transcribed TFBS, and the decrease in repair rate in XP-C cells is apparent for TFs bound to either template or non-template strand (Extended Data Fig. 7). This result demonstrates that the decrease in NER caused by bound TFs results from impairment of both NER pathways.

NER specifically recognizes and repairs other DNA lesions beside those induced by UV light, such as DNA adducts caused by smoking-related carcinogens (for example, benzo[*a*]pyrene diol epoxide)[26]. We therefore hypothesized that the observations made in melanomas could be extended to tobacco-related tumours. We observed higher mutation rates at TFBS in lung adenocarcinomas and lung squamous cell carcinomas, in particular for C > A variants, which correspond to the mutations caused by tobacco smoking[11] (Extended Data Fig. 8). In contrast, no increment of the mutation rate in TFBS was observed in colon adenocarcinomas, where NER activity is not expected to play a major role in shaping the mutational process, and only modest increments are detected in other tumour types (Extended Data Fig. 9).

Two previous studies have described abnormal mutation rates in connection with a group of DNA bound TFs in yeast[17] and CTCF/cohesin sites in a subset of colorectal tumours[18]. However, in contrast to our results, in neither of these studies was the increased mutation rate caused by impairment of NER resulting from bound proteins. In the former, elevated mutation rate at specific TFBS was related to polymerase δ mediated displacement of polymerase α synthesized DNA during replication. In the latter study, higher mutations at CTCF/cohesin sites in a subset of colorectal tumours was attributed to challenged DNA replication under aberrant conditions. Also, earlier biochemical studies[7,8] focusing on two short individual promoter DNA regions observed that the repair of CPDs in TFBS was slower than those in unbound DNA and the authors speculated on the potential effect this could have on the mutation rates at such sites. The interplay between different rates in the generation of UV-induced damage, its effect on DNA–protein binding, and the rate of the repair of lesions on the mutation rate at the local level in promoter regions was nevertheless not clear[16,27]. Here, for the first time (to our knowledge), we have uncovered

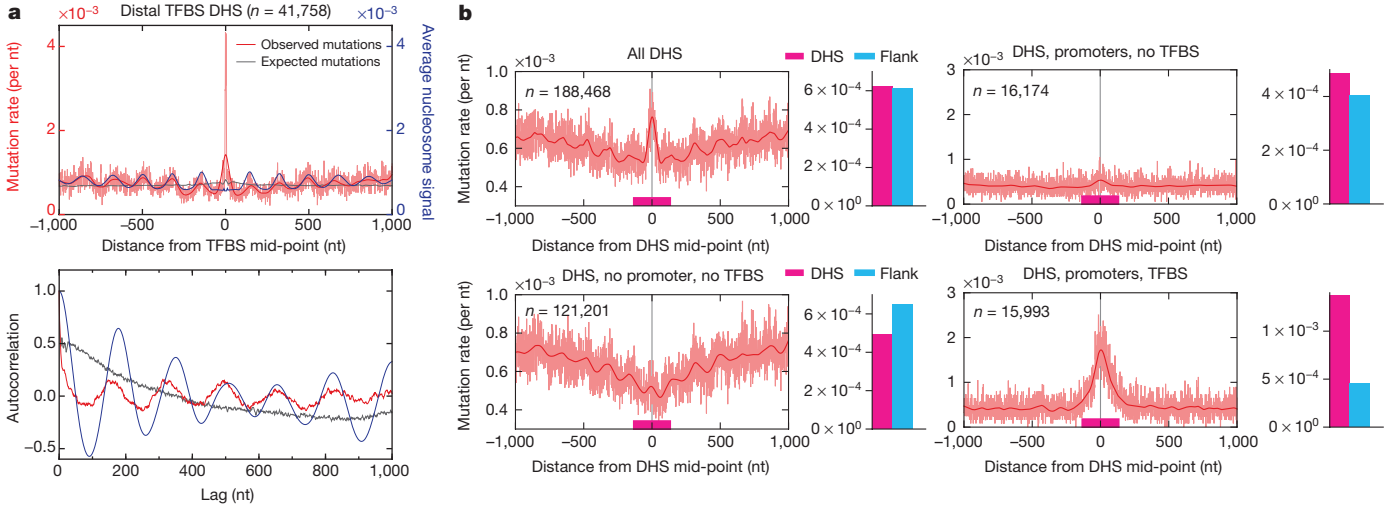


Figure 2 | Mutation rate at distal TFBS and DHS sites. a, Mutation rate in distal TFBS, which are 5 kb away from transcription start sites. Similar to proximal TFBS as shown in Fig. 1a, the mutation rate is elevated at the centre of core TFBS compared to the flanking sites. In addition, periodic peaks of mutation rate in the flanking regions of binding sites correlate well with the nucleosome positioning (blue line). This is further supported by the autocorrelation analysis (bottom panel) that shows the periodic peaks are observed at a distance of ~170 bp, which coincides well with the size of the DNA being wrapped around nucleosomes (~146 bp) and the linker DNA. The periodicity in mutation rate is not explained by sequence context; black lines show the expected mutation rate per position when distributing all observed mutations in the region according to the probability of mutations with different trinucleotide contexts. **b**, Mutation rate centred in DHS in melanocytes is shown (top panel). In the subset of DHS outside promoter regions which do not contain sequences of any overlapping TFBS (DHS, no promoter, no TFBS), the peak of mutation rate disappears and only a valley is observed. In the subset of DHS regions in promoters overlapping TFBS (DHS, promoters, TFBS) there is a 2.5-fold change increase ($P < 2.2 \times 10^{-16}$) of mutation rate in the DHS compared to the flanking region. In contrast, only a modest increase (FC = 1.18, $P = 5.3 \times 10^{-10}$) is observed in the subset of DHS regions in promoters that do not contain sequences of any overlapping TFBS (DHS, promoter, no TFBS), probably due to remaining TFBS not detected by our analysis. The actual mutation rate values are shown in light red and the best-fit spline is shown in dark red. The zero coordinate in the *x* axis corresponds to the DHS peak mid-point, and the magenta line above it represents the average size of DHS (~150 nt). The bar plot at the right of each panel compares the mutation rate in the DHS and the flank for each group of regions.

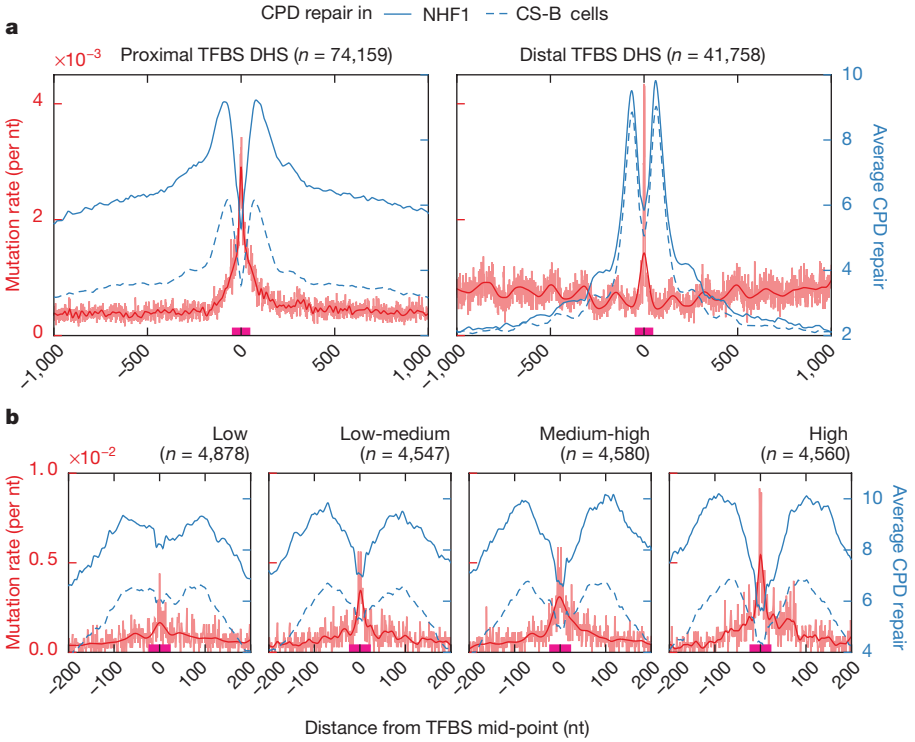


Figure 3 | Regions around TFBS show a decrease in nucleotide excision repair. a, Mutation rate around TFBS is plotted (red line) alongside the average repair of UV-light induced DNA damage, cyclobutane pyrimidine dimers (CPD), in wild-type (NHF1) and CS-B mutant cell lines (blue lines). A sharp decrease in nucleotide excision repair is evident at the core TFBS both proximal and distal. **b**, The level of nucleotide excision repair (and the resulting mutation rate) in TFBS correlates with the strength of the binding of the transcription factor to its site. The binding sites were classified into four quartiles (low to high) using the ChIP-seq read coverage that reflects the strength of binding or occupancy (as in ref. 17). The binding sites in the 'high' quartile (last panel) show higher mutation rates at the centre (correlating with the lower repair) compared to the 'low' quartile (first panel). The zero coordinate in the *x* axis corresponds to the TFBS mid-point, and the magenta line above it represents the average size of TFBS.

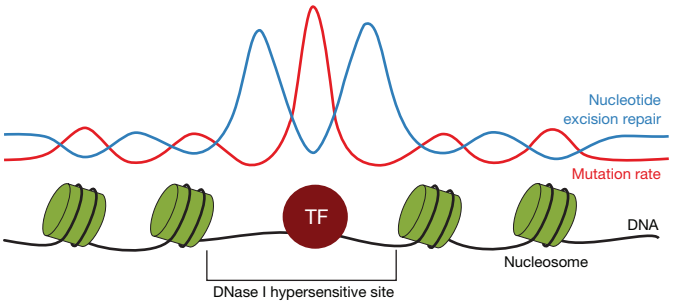


Figure 4 | Model showing the mutation rate and repair rate in TFBS and nucleosome sites. The model shows that the accessibility of the DNA to the nucleotide excision repair (NER) machinery directly determines the distribution of mutational density at the nucleotide scale. Lower NER activity is observed at the TFBS bound region (within DHS region) and the nucleosome positions in the flank, compared to the nucleosome free regions (DHS and linkers). This decrease in NER activity is the cause of the observed high mutation rate in transcription factor and nucleosome bound regions.

the genome-wide elevated mutation rate at TFBS of UV-exposed cells and established a causative link between the impairment of NER and proteins bound to the DNA.

Our results demonstrate that the accessibility of the DNA to the NER machinery directly affects the distribution of mutational density at the nucleotide scale. The increased repair in freely accessible, nucleosome-free, DNA around TFBS, and the decline in repair efficiency exactly at TFBS produces a lower mutation rate at the periphery of DHS sites and higher mutation rate at their centre (Fig. 4). Moreover, periodic signals of higher mutation rate and lower NER in close chromatin regions coincide with nucleosome occupancy, suggesting that nucleosomes produce the same type of impairment to NER.

These findings indicate that most mutations in TFBS accumulate due to faulty repair at these sites. Therefore, methods designed to identify potential somatic driver mutations, in non-coding regions, which typically exploit the mutational patterns of genomic elements must construct models of the background mutation rate that accurately take into account this fact. The causative link between impairment of NER and proteins bound to the DNA have strong implications for our basic understanding of how the mechanisms of DNA repair in human cells shape their mutational profile, as well as for the study of tumour evolution and cancer-associated somatic mutations.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 14 October 2015; accepted 15 March 2016.

1. Yates, L. R. & Campbell, P. J. Evolution of the cancer genome. *Nature Rev. Genet.* **13,** 795–806 (2012).
2. Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488,** 504–507 (2012).
3. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).
4. Polak, P. *et al.* Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. *Nature Biotechnol.* **32,** 71–75 (2014).
5. Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518,** 360–364 (2015).
6. Hu, J., Adar, S., Selby, C. P., Lieb, J. D. & Sancar, A. Genome-wide analysis of human global and transcription-coupled excision repair of UV damage at single-nucleotide resolution. *Genes Dev.* **29,** 948–960 (2015).
7. Gao, S., Drouin, R. & Holmquist, G. P. DNA repair rates mapped along the human PGK1 gene at nucleotide resolution. *Science* **263,** 1438–1440 (1994).
8. Conconi, A., Liu, X., Koriazova, L., Ackerman, E. J. & Smerdon, M. J. Tight correlation between inhibition of DNA repair *in vitro* and transcription factor IIIA binding in a 5S ribosomal RNA gene. *EMBO J.* **18,** 1387–1396 (1999).
9. The International Cancer Genome Consortium International network of cancer genome projects. *Nature* **464,** 993–998 (2010).
10. The Cancer Genome Atlas Research Network The Cancer Genome Atlas pan-cancer analysis project. *Nature Genet.* **45,** 1113–1120 (2013).
11. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).
12. Supek, F. & Lehner, B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* **521,** 81–84 (2015).
13. Hara, R., Mo, J. & Sancar, A. DNA damage in the nucleosome core is refractory to repair by human excision nuclease. *Mol. Cell. Biol.* **20,** 9173–9181 (2000).
14. Yazdi, P. G. *et al.* Increasing nucleosome occupancy is correlated with an increasing mutation rate so long as dna repair machinery is intact. *PLoS ONE* **10,** e0136574 (2015).
15. Tolstorukov, M. Y., Volfovsky, N., Stephens, R. M. & Park, P. J. Impact of chromatin structure on sequence variability in the human genome. *Nature Struct. Mol. Biol.* **18,** 510–515 (2011).
16. Tornaletti, S. & Pfeifer, G. P. UV damage and repair mechanisms in mammalian cells. *Bioessays* **18,** 221–228 (1996).
17. Reijns, M. A. *et al.* Lagging-strand replication shapes the mutational landscape of the genome. *Nature* **518,** 502–506 (2015).
18. Katainen, R. *et al.* CTCF/cohesin-binding sites are frequently mutated in cancer. *Nature Genet.* **47,** 818–821 (2015).
19. Fredriksson, N. J., Ny, L., Nilsson, J. A. & Larsson, E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nature Genet.* **46,** 1258–1263 (2014).
20. Martincorena, I. *et al.* High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348,** 880–886 (2015).
21. Marteijn, J. A., Lans, H., Vermeulen, W. & Hoeijmakers, J. H. J. Understanding nucleotide excision repair and its roles in cancer and ageing. *Nature Rev. Mol. Cell Biol.* **15,** 465–481 (2014).
22. Tornaletti, S. & Pfeifer, G. P. UV light as a footprinting agent: modulation of UV-induced DNA damage by transcription factors bound at the promoters of three human genes. *J. Mol. Biol.* **249,** 714–728 (1995).
23. Gale, J. M., Nissen, K. A. & Smerdon, M. J. UV-induced formation of pyrimidine dimers in nucleosome core DNA is strongly modulated with a period of 10.3 bases. *Proc. Natl Acad. Sci. USA* **84,** 6644–6648 (1987).
24. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
25. Goodman, M. F. & Woodgate, R. Translesion DNA polymerases. *Cold Spring Harb. Perspect. Biol.* **5,** a010363 (2013).
26. Nouspikel, T. DNA repair in mammalian cells. *Cell. Mol. Life Sci.* **66,** 994–1009 (2009).
27. Wyrick, J. J. & Roberts, S. A. Genomic approaches to DNA repair and mutagenesis. *DNA Repair (Amst.)* **36,** 146–155 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements We acknowledge funding from the Spanish Ministry of Economy and Competitiveness (grant number SAF2012-36199), the Marató de TV3 Foundation, and the Spanish National Institute of Bioinformatics (INB). R.S. is supported by an EMBO Long-Term Fellowship (ALTF 568-2014) co-funded by the European Commission (EMBOCOFUND2012, GA-2012-600394) support from Marie Curie Actions. A.G.-P. is supported by a Ramón y Cajal contract (RYC-2013-14554).

Author Contributions N.L.-B. conceived and supervised the study. N.L.-B. and R.S. designed the analyses. R.S. performed the analyses with contributions from L.M. and J.D.-P. All authors participated in the discussion of the results. N.L.-B., A.G.-P. and R.S. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to N.L.-B. (nuria.lopez@upf.edu).

METHODS

Mutation data. Whole-genome somatic mutations of 38 skin cutaneous melanomas (SKCM), 46 lung adenocarcinomas (LUAD), 45 lung squamous cell carcinomas (LUSC), 42 colorectal adenocarcinomas (CRC), 96 breast cancer (BRCA), 21 bladder cancer (BLCA) and 27 head and neck squamous cell carcinoma (HNSC) identified by TCGA were obtained from Fredriksson *et al.*, 2014[19]. As suggested by the authors of that paper, we considered in our analyses only single nucleotide substitutions with a minimum variant frequency of 0.2 and which do not overlap dbSNP entries (v138). The total number of mutations of each cancer type passing these thresholds is listed in Extended Data Table 1. We separated CRC samples into two groups: hypermutated (with mutations of the DNA polymerase epsilon (POL-E) gene; $n = 8$ samples) and hypomutated (the rest; $n = 34$ samples). In addition, mutations detected across the whole genome of a normal human skin sample were obtained from ref. 20 and treated as an independent data set.

Genomic elements. The genomic coordinates of transcription factor binding sites (TFBS), that is, TF motif match under ChIP-seq peak regions, were obtained from ENCODE[24]. These comprised the binding sites of 109 transcription factors (TFs) as used in ref. 28. We also obtained from ENCODE predicted binding sites of 52 transcription factors which are not supported by ChIP-seq peaks (termed unbound TFBS). In addition, we obtained the binding sites of 32 TFs used in ref. 17. We treated the latter as an independent data set, and following the authors of the original paper[17], we clustered the TFBS into quartiles according to the binding strength or occupancy of the TFs to their sites—quantified through ChIP-seq read coverage.

As promoters, we considered the DNA sequences up to 2.5kb upstream of transcription start sites (TSS) of all protein coding genes in GENCODE[29] (v19). Promoter regions overlapping coding sequences (CDS) or untranslated regions (UTRs) were excluded. We classified TFBS as either proximal (overlapping these upstream promoters) or distal (those located in intergenic regions, with no annotated TSS (as per GENCODE v19) within 5kb distance on either side). A third group of TFBS was composed of those located downstream TSS (between +200 bp and +500 bp) and which do not overlap with the upstream 2.5 kb promoter regions—that is, TFBS in transcribed regions.

All TFBS overlapping DNase I hypersensitive sites (DHS) identified by the Epigenome Roadmap Project[30] in primary cell types most closely matching the cell of origin of each tumour type (see below) were considered active. We considered only DHS sites identified by the Hotspot algorithm (narrowPeaks in FDR 1%), which are typically 150 nt long. For each cancer type, the matching primary cell type was selected based on a recent study[5] (Extended Data Table 1). We chose the DHS from primary cell types (from the Epigenome Roadmap Project) instead of cell lines (from ENCODE), because the chromatin features of the cell of origin of a tumour has been shown to correlate better with its mutation profile than that of matched cancer cell lines[5]. However, we selected the TFBS detected by ENCODE in cell lines (see above) due to the lack of TF binding site annotations in primary cells analysed by the Epigenome Roadmap Project[30]. Only for two cancer types (BLCA and HNSC) the closest matching primary cell types is not available in Epigenome Roadmap project, and in those cases we used the DHS from ENCODE[24] (Extended Data Table 1). As the TFBS from ENCODE cover a limited number of TFs ($n = 109$), we employed the PIQ[31] algorithm to predict TFBS for 1,316 TFs using the DNase profiles from melonocytes[30]. This resulted in 2,553,927 high-confidence binding sites (with purity score >0.8) for 1,284 TFs in DHS across the entire genome. We treated these predicted TFBS as an independent data set and used them in the DHS-centred analysis (in Fig. 2b and Extended Data Figs 3 and 6).

We then classified the TFBS in the samples of each tumour type as active or inactive based on their overlap, or lack thereof, with DHS regions (minimum 1 bp) of the matched primary cell type. Unbound TFBS (see above), which do not overlap with TF peaks or DHS regions, were considered as inactive TFBS and used as negative control to compare with the active TFBS (in Extended Data Fig. 1). All genomic coordinates of TFBS used in this study as part of any aforementioned category are available at http://bg.upf.edu/tfbs.

Mutation rate estimation. In order to compare the mutation rate in TFBS to their neighbouring regions, we considered flanking stretches of 1,000 nucleotides at both sides of the TFBS mid-point. To exclude regions that could bias the mutation rate analyses, before mapping the somatic mutations to these selected 2001 nt windows, we filtered out: any regions overlapping (a) coding sequences, and (b) UCSC Browser blacklisted regions, often misaligned to sites in the reference assembly, (Duke and DAC) and low unique mappability of sequencing reads ("CRG Alignability 36' Track"[32], score <1) (http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeMapability). In addition, regions that overlap other TFBS within flanking regions (immediately upstream or downstream the TFBS) were excluded. The resulting filtered windows of each TFBS were then aligned (taking as reference the TFBS centres), and the mutation rate of every column i within the window was calculated as the total number of mutations mapped to nucleotides in column i divided by the total number of nucleotides observed in column i (after filtering). We computed this mutation rate for each TF separately, as well as globally for all TFs. In the latter case, before the calculation, we removed any repeated chromosomal positions (from different TFs) observed in a column.

In the case of the analysis centre on DHS, we considered flanking stretches of 1,000 nucleotides at both sides from DHS peak centre and followed the same steps mentioned above to filter mutations and to compute the mutation rate.

Background mutation rate estimation. In order to check if the mutation rate observed at each position was expected due to the local sequence context, we randomly introduced the same number of mutations observed at each window following the probability of occurrence of each mutation according to its tri-nucleotide context. We computed the probability of occurrence of all possible 96 tri-nucleotide changes in each cancer type based on the total number of observed mutations in all its samples. We also computed separate probabilities of occurrence of all 96 tri-nucleotide in active and inactive TFBS from the mutations observed in each category. The mutation rate of each randomly generated set of changes, was computed for each column as explained above. This procedure was repeated 1,000 times to compute the mean random mutation rate of every column in the motif.

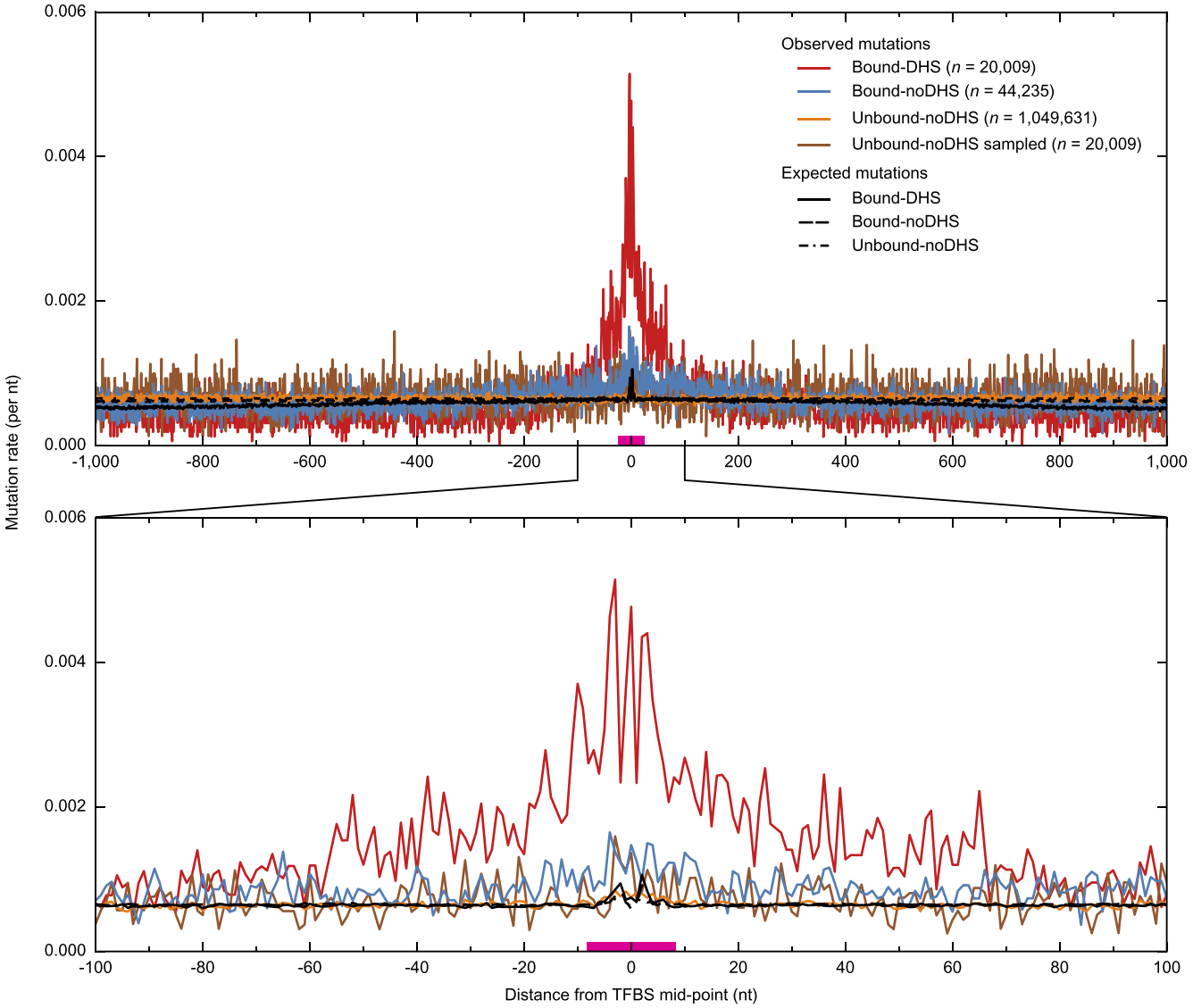
Enrichment analysis. To test the enrichment for mutations on TFBS and DHS sites compared to the immediate flanking region, we compared the ratio of the total number of mutations to the total number of nucleotide positions within the TFBS region (−15 to +15 nt) or DHS sites (−75 to +75 nt) and that of the flanking region (16 to 1,000 nt or 76 to 1,000 nt on either side respectively) using a chi-squared test. We performed this test for all transcription factors and for each individual tumour, and corrected the resulting P values for multiple-testing using the Benjamini–Hochberg procedure[33]. In addition, we computed the fold change of mutation rates through the expected frequencies obtained from chi-squared tests. Both, the fold change and adjusted P values are shown in Fig. 1b, c.

Nucleotide excision repair data. The genome-wide maps of nucleotide excision repair of two types of UV-induced damage, CPD and 6–4PP, available for three different cell lines (i) wild-type NHF1 skin fibroblasts, (ii) XP-C mutants, lacking the global repair mechanism, and (iii) CS-B mutants lacking transcription-coupled repair, were obtained from ref. 6. The data set contains normalized read counts for fixed steps of 25 bp across the genome, for the forward and reverse strands separately. We kept these for our analyses and also generated strand independent data as the average of normalized read counts from both strands for every nucleotide position. These average read counts were mapped to the TFBS centred windows (2,001 bp), filtered and aligned to the TFBS mid-point as described above. We computed the average repair rate for each column i of these windows as the total number of average read counts mapped to the nucleotides in the column i divided by the total number of nucleotides in the column i, as described above for the mutation rate.

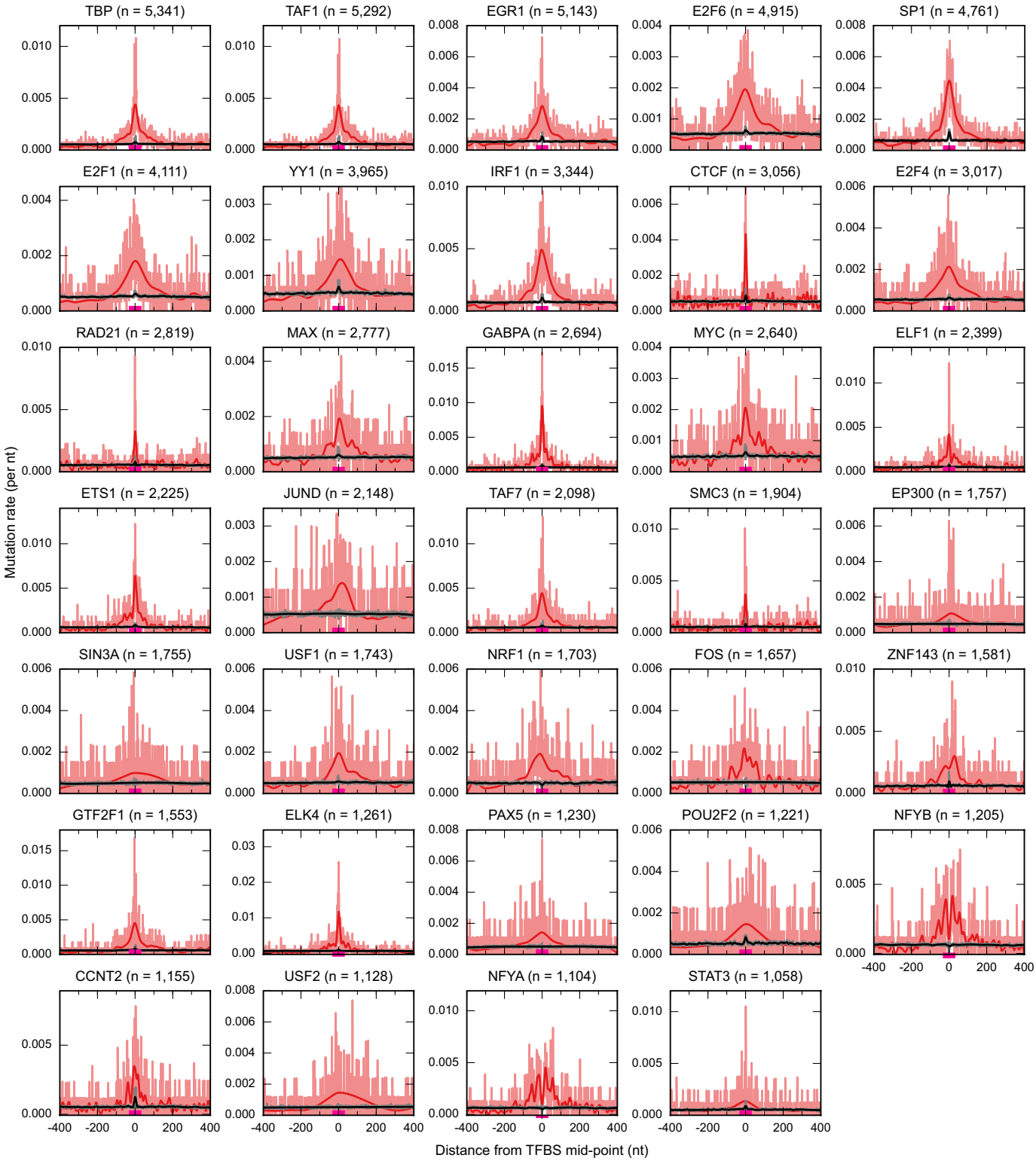
Nucleosome signals. Genome-wide nucleosome positioning signals (density graph) of ENCODE cell line GM12878 (lymphoblastoid cell line) were downloaded via the UCSC genome browser (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeSydhNsome/). We then mapped them to the TFBS centred windows, and similar to mutation and repair rates, we computed the average signal per column i of the window as the sum of signal values mapped to the nucleotides in column i divided by the total number of nucleotides in column i.

Computational and statistical tools. BEDTools utilities[34] were used to carry out operations as extensions or overlaps in the various analyses of genomic features (TFBS/DHS), as well as to map somatic mutations to genomic features. All curve fittings shown in figures (best-fit spline) were performed using the smooth.spline function from R[35] (v3.0). The auto-correlation was performed using the *acf* function from statsmodels python package (http://statsmodels.sourceforge.net/). The code used to run the analyses and generate figures is available from http://bg.upf.edu/tfbs.

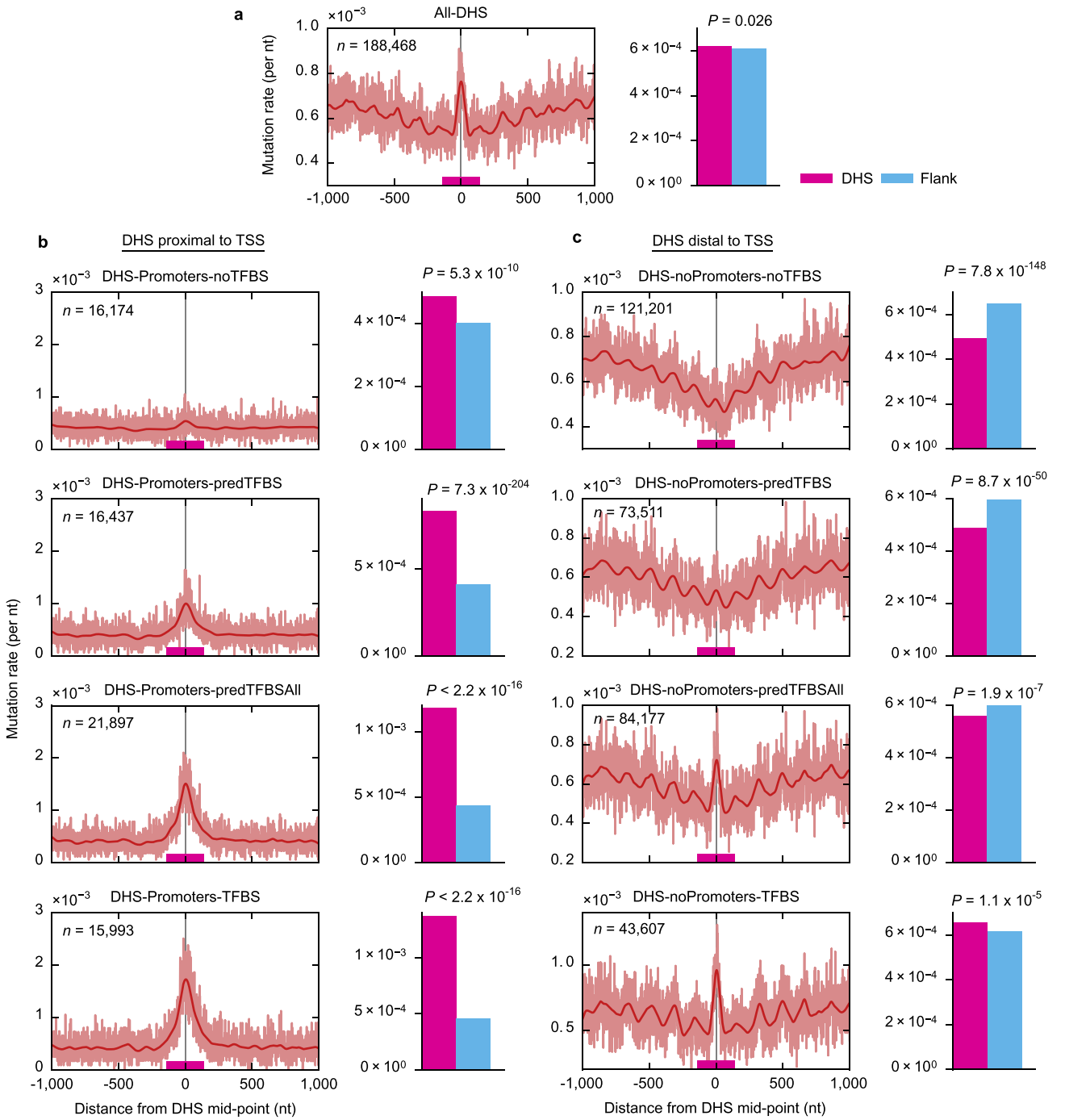
28. Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342,** 1235587 (2013).
29. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22,** 1760–1774 (2012).
30. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).
31. Sherwood, R. I. *et al.* Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nature Biotechnol.* **32,** 171–178 (2014).
32. Derrien, T. *et al.* Fast computation and applications of genome mappability. *PLoS ONE* **7,** e30377 (2012).
33. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57,** 289–300 (1995).
34. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
35. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2014).
36. Haradhvala, N. J. *et al.* Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. *Cell* **164,** 538–549 (2016).



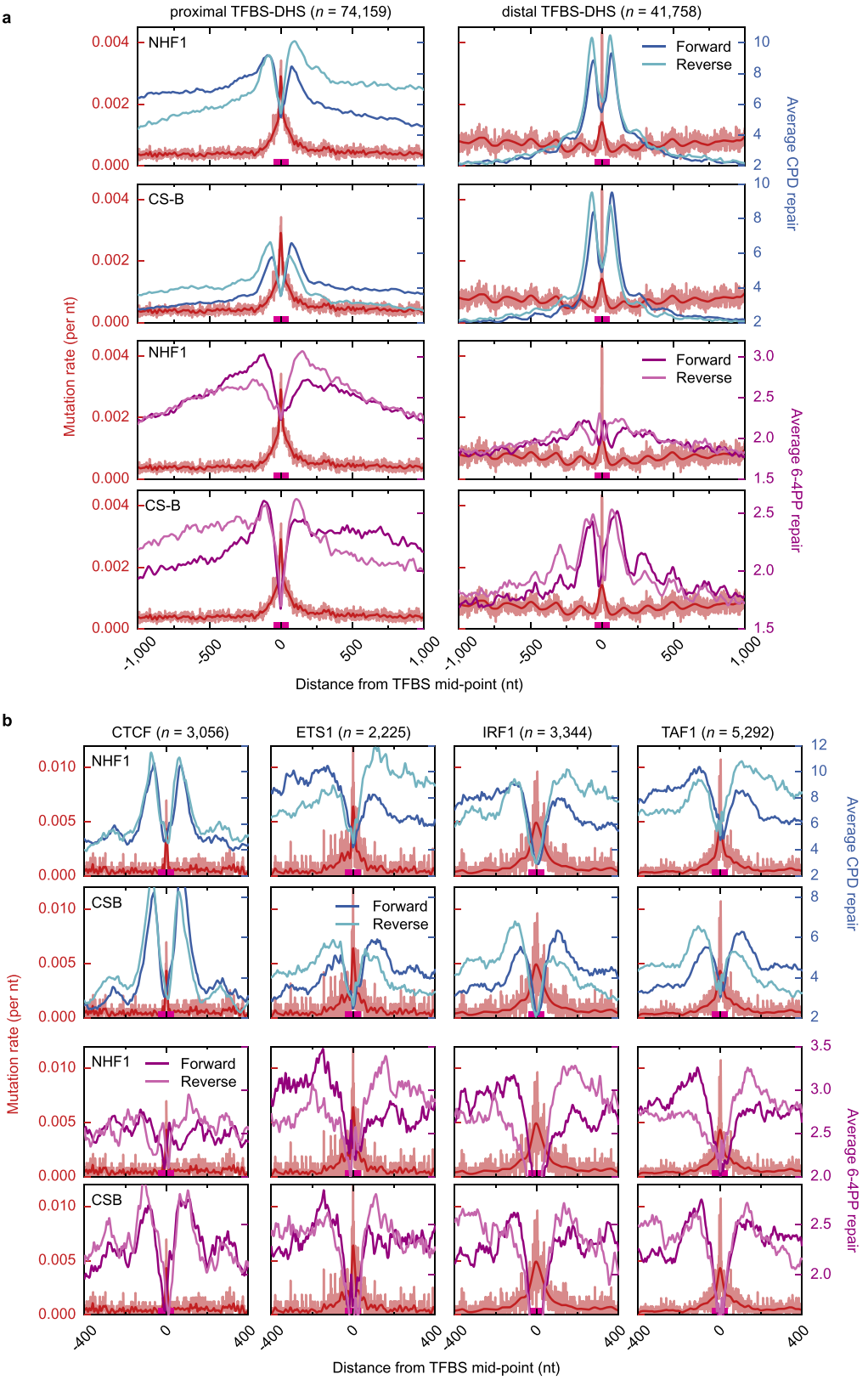
Extended Data Figure 1 | Higher mutation rate at bound TFBS compared to unbound TFBS in melanoma. The mutation rate is higher in active TFBS (bound by their TF and overlapping DHS; bound-DHS, red line) compared to: (i) inactive TFBS (not overlapping any DHS; bound-noDHS, blue line); and (ii) unbound inactive TFBS (not bound by TF and not overlapping any DHS; unbound-noDHS, orange line). The binding sites considered here correspond to the subset of TFs ($n = 58$) for which both the bound and unbound motif predictions are available from the ENCODE integrative analysis[24]. For comparison purposes, we sampled an equal number of unbound-noDHS TFBS (unbound-noDHS sampled, brown line) as in the set of bound-DHS, and confirmed that the mutation rate is still higher in the bound TFBS. The background mutation rates of each group are represented as black lines. The zero coordinate in the *x* axis corresponds to the TFBS mid-point, and the magenta line above it represents the average size of TFBS.



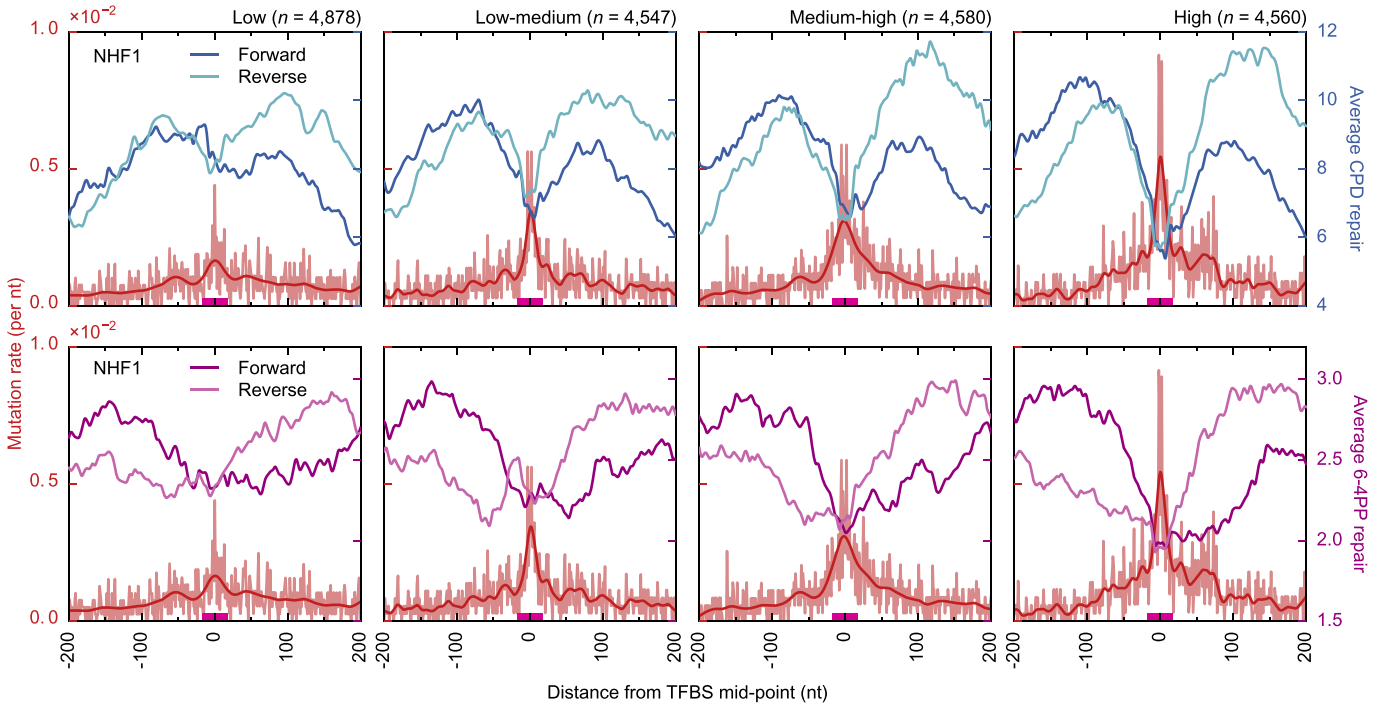
Extended Data Figure 2 | Elevated mutation rate at the binding sites of individual transcription factors in melanoma. Here, we show the mutation rate of the TFBS of all TFs with at least 1,000 binding sites overlapping melanocytes DHS. The observed mutation rate is shown in red (light colour in the background corresponds to the actual data points, and the thick solid line on top is the best-fit spline), while the background mutation rate is represented by the black line. The zero coordinate in the *x* axis corresponds to the TFBS mid-point, and the magenta line above it represents the average size of TFBS.



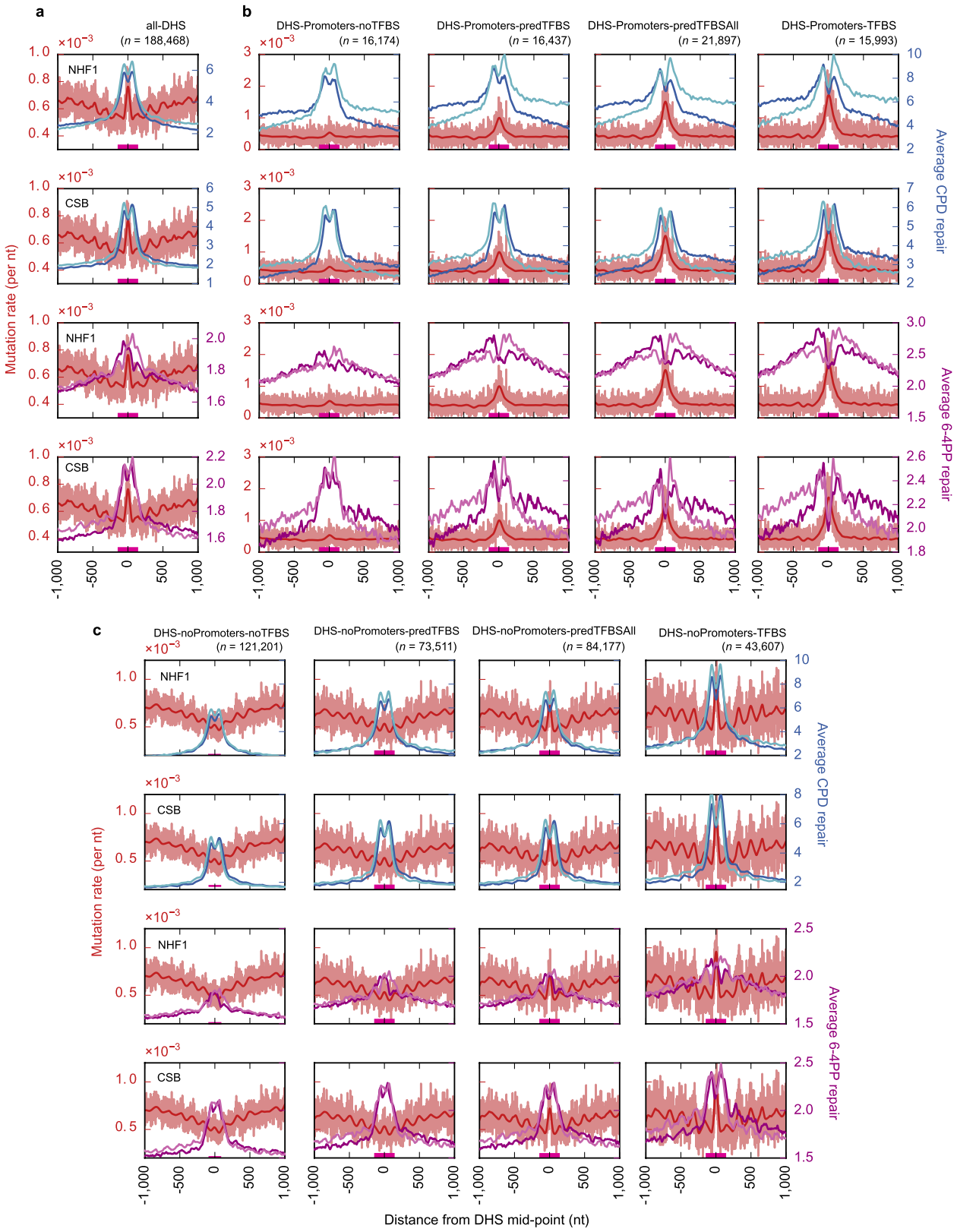
Extended Data Figure 3 | Mutation rate at DHS sites. a–c, Mutation rate centred in DHS sites in melanomas is shown for all DHS genome wide (**a**), a subset of DHS regions overlapping promoters (2.5kb from TSS) (**b**) and DHS regions outside promoters (**c**). Within **b** and **c**, the first row shows the mutation rate in regions that do not contain sequences of any overlapping TFBS (noTFBS), neither predicted TFBS (from PIQ[31], corresponding to 1,284 different motifs) or known TFBS (mapped from ENCODE[28] ChIP-seq analysis, corresponding to 109 TFs). The second row contains only predicted TFBS (predTFBS), removing any sequences that overlap the known TFBS. The third row contains the subset of sequences that overlap with all predicted TFBS, without removing the known ones (predTFBSall). The last row contains the subset of sequences with known TFBS. The barplot at the right of each panel compares the mutation rate in the DHS and the flank for each group of regions, and the *P* value (from chi-square test) shows the enrichment of mutation rate between two groups. The increase in predicted TFBS is, as expected, lower than that observed within the TFBS mapped by ENCODE (DHS-promoter-TFBS), reflecting the lower precision in the mapping of the predictions compared to mapping by ChIP-seq. The zero coordinate in the *x* axis corresponds to the DHS peak mid-point, and the magenta line above it represents the average size of DHS (~150 nt).



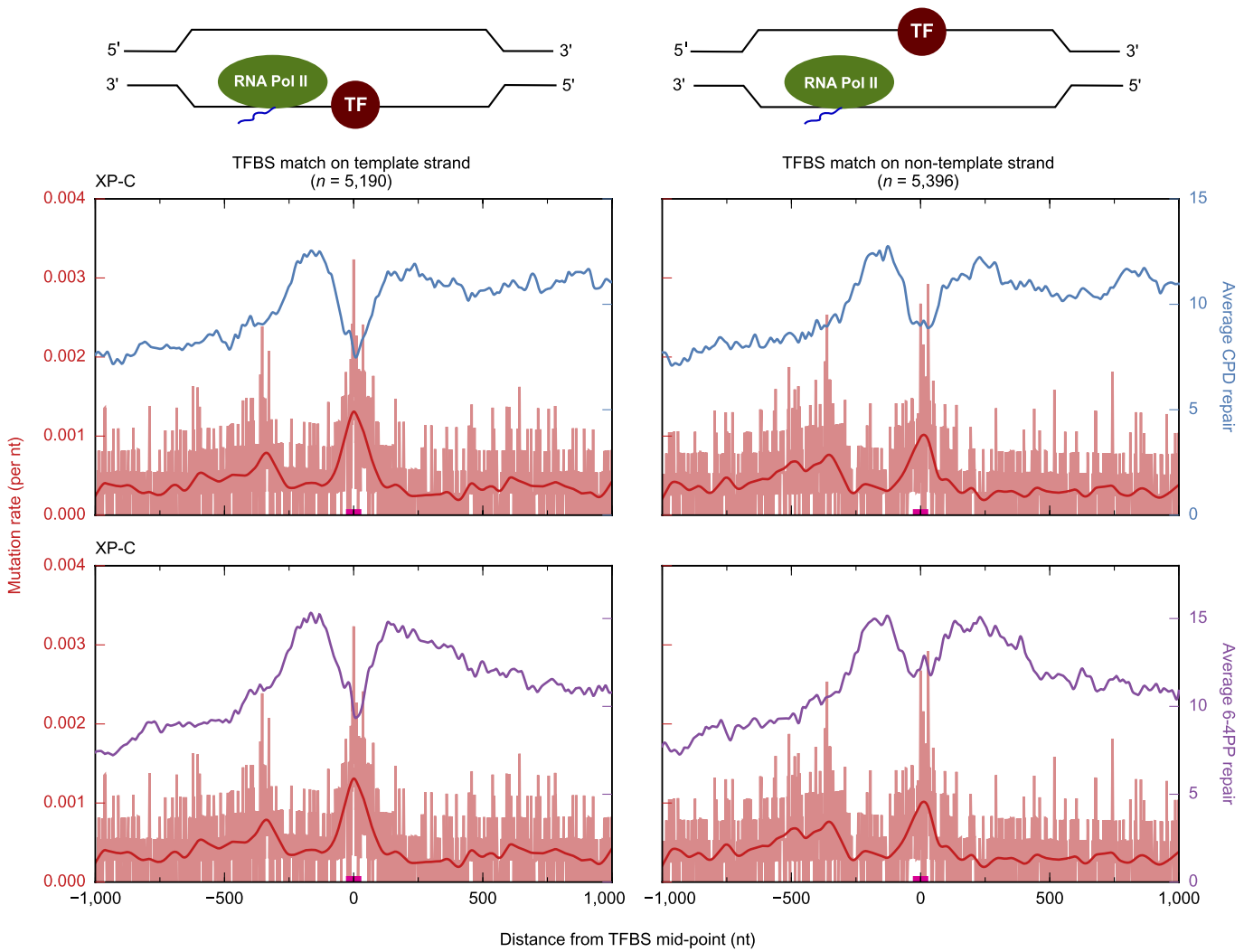
Extended Data Figure 4 | Regions around TFBS show a decrease in nucleotide excision repair. Mutation rate around TFBS plotted alongside the average repair of two types of UV-light induced DNA damage—CPD and 6–4PP in wild-type NHF1 cell line of skin fibroblasts and the CS-B mutant cell line for proximal (left column) and distal (right column) TFBS in **a**. Also, a lower level of nucleotide excision repair is observed at the binding sites of individual transcription factors. For example, the results for CTCF, ETS1, IRF1 and TAF1 are shown in **b**. In both **a** and **b**, the observed mutation rate is shown in red (light colour in the background corresponds to the actual data points, and the thick solid line on top is the best-fit spline). The two top rows show the CPD repair on NHF1 and CS-B cells, respectively and the two bottom rows show the 6–4PP repair on NHF1 and CS-B cells, respectively. Here the average repair levels are shown separately for the forward and reverse strands of the genome (as obtained from ref. 6).



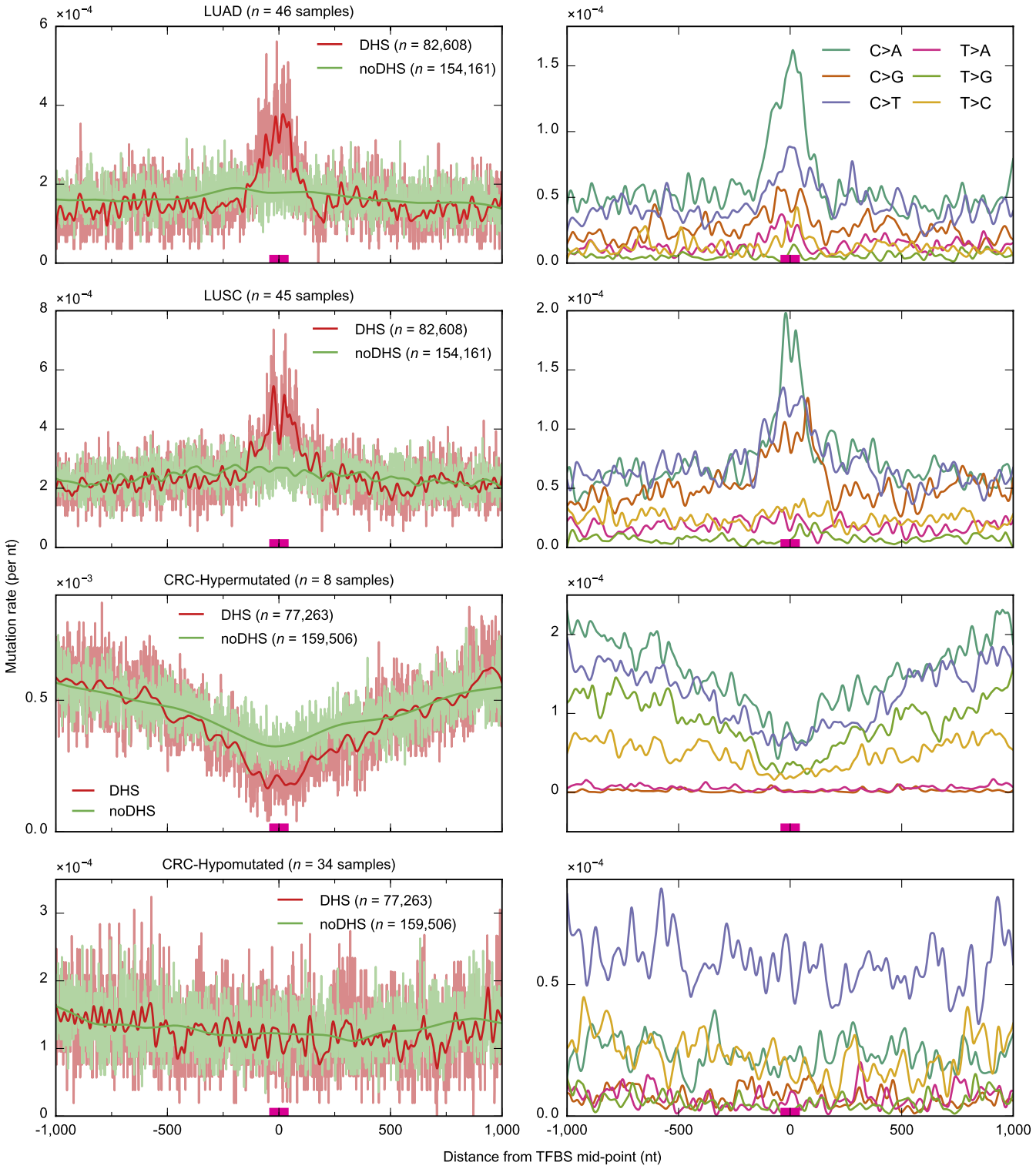
Extended Data Figure 5 | The level of nucleotide excision repair, and the resulting mutation rate in TFBS correlate with the strength of the binding signal of transcription factors to their sites. Regions around TFBS sites were obtained from ref. 17. As in ref. 17, the binding sites were classified into four quartiles (low to high) using the ChIP-seq read coverage that reflects the strength of binding or occupancy. The binding sites in the 'high' quartile (fourth column) tend to bear higher mutation rates at the centre (correlating with lower repair) compared to the 'low' quartile (first column). The nucleotide excision repairs of two photoproducts (CPD and 6–4PP) shown here are from NHF1 wild-type cell line. Average repair levels are shown separately for the forward and reverse strands of the genome (as obtained from ref. 6).



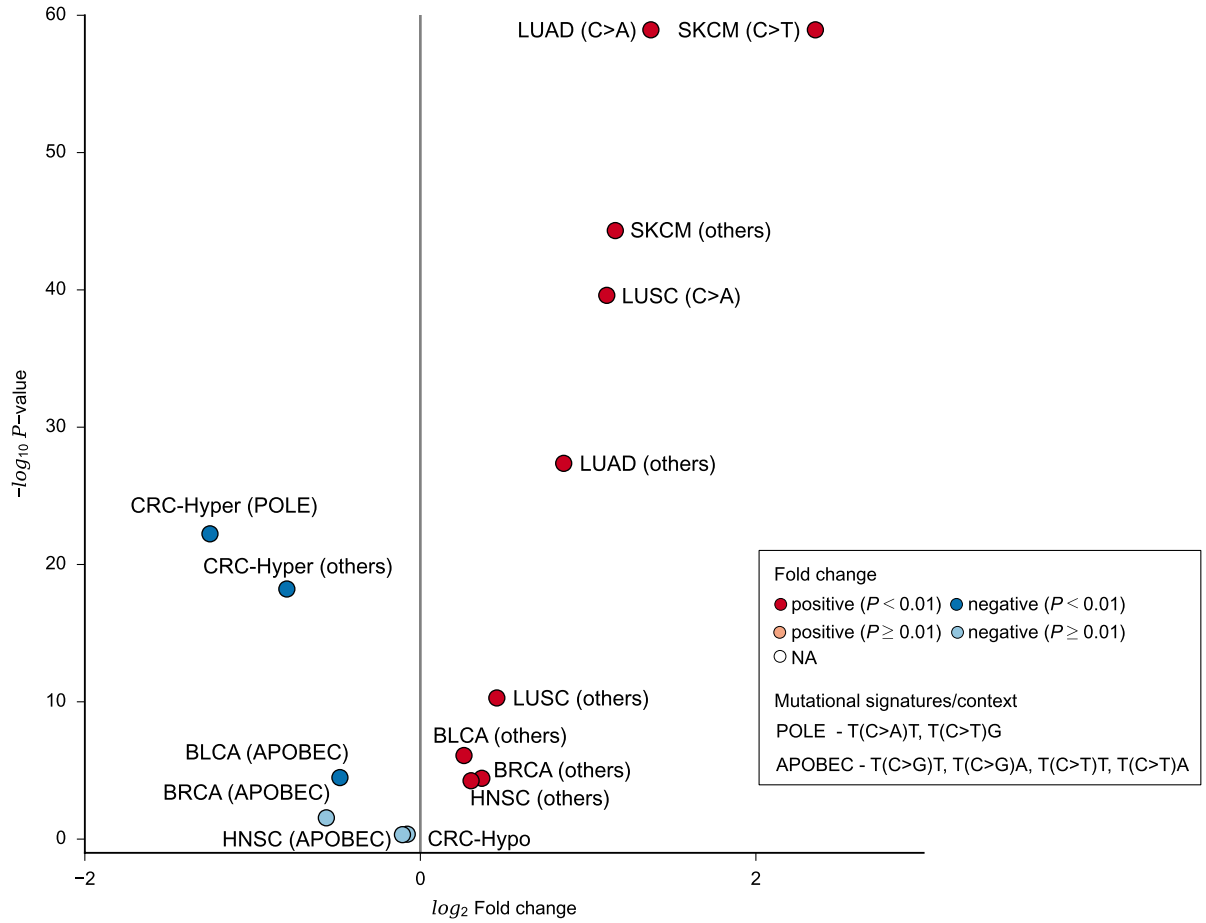
Extended Data Figure 6 | Nucleotide excision repair and mutation rate in DHS centred regions. a–**c**, The distribution of nucleotide excision repair, for the two types of UV-light induced DNA damages, is shown for all DHS genome-wide (**a**), DHS regions overlapping promoters (2.5 kb from TSS) (**b**) and DHS regions outside promoters (**c**). Within **b** and **c** the first column shows the mutation rate in regions that do not contain sequences of any overlapping TFBS (noTFBS), neither predicted TFBS (from PIQ[31], corresponding to 1,284 different motifs) or known TFBS (mapped from ENCODE[28] ChIP-seq analysis, corresponding to 109 TFs). The second column contains only predicted TFBS (predTFBS), removing any sequences that overlap the known TFBS. The third column contains the subset of sequences that overlap with all predicted TFBS, without removing the known ones (predTFBSall). The last column contains the subset of sequences with known TFBS. The two top rows in **a**, **b** and **c** show the CPD repair on NHF1 and CS-B cells, respectively and the two bottom rows show the 6–4PP repair on NHF1 and CS-B cells, respectively. Here average repair levels are shown separately for the forward and reverse strands of the genome (as obtained from ref. 6). The zero coordinate in the *x* axis corresponds to the DHS peak mid-point, and the magenta line above it represents the average size of DHS (~150 nt).



Extended Data Figure 7 | Transcription coupled-repair is impaired at active TFBS. To carry out this analysis, TFBS overlapping transcribed regions (located 200–500 bp downstream of TSS) were centred at the TFBS mid-point. We plotted the mutation and repair rates of UV induced damages (CPD and 6–4PP) in XP-C cells, which possess only transcription-coupled repair capability. TFBS in either strand were separated: those in the template strand of the gene are shown in the left panel, while those in the non-template strand are presented in the right panel. All TFBS and their flanking regions are shown in the same orientation (5′ to 3′). This result shows that TF binding to both strands results in lower transcription-coupled NER activity.



Extended Data Figure 8 | Mutation rate around TFBS in other cancer types. Mutation rates around TFBS of promoter regions of lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), and colorectal cancer (CRC) are shown. CRC samples are separated into two groups, those with missense mutations of the DNA polymerase epsilon (POL-E) gene or hypermutated ($n = 8$ samples) and the rest or hypomutated ($n = 34$ samples). In the left column, the mutation rate is shown for active TFBS that overlap DHS sites (red line) and inactive TFBS that do not overlap DHS (green line). The right column graphs present the mutation rate of six different changes separately in active TFBS. In lung cancers (LUAD and LUSC), C > A changes, caused by tobacco carcinogens, contributes more to the elevated mutation rate, which indicates that NER activity is lower at these active TFBS.



Extended Data Figure 9 | Mutation enrichment around TFBS across cancer types. Overrepresentation of mutations at TFBS as compared to their immediate flanking regions for different cancer types and mutational signatures. The mutational process/signatures specific to each cancer type are defined as in ref. 36: UV-light associated signature (C>T) in melanoma (SKCM), tobacco smoking associated signature (C>A) in lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC), mutated POL-E associated signatures (T(C>A)T, T(C>T)G) in colorectal samples, and APOBEC associated mutational signature (T(C>G)T, T(C>G)A, T(C>T)T, T(C>T)A) in breast (BRCA), bladder (BLCA) and head and neck squamous cell carcinomas (HNSC). Mutations in each sample that don't follow the aforementioned mutational signatures are grouped into one class (referred to as 'others') for each cancer type. The $\log_2$ fold change in the *x* axis represents how much higher (positive fold change) or lower (negative fold change) than the expected the observed mutation rate in TFBS is; the corresponding significance value (derived from a chi-square test) is shown on the *y* axis for each cancer type-signature combination. These results show that the only tumour samples with mutations clearly overrepresented at TFBS are lung carcinomas and melanomas. In both cases it is the predominant mutational signature, induced by the external mutagenic agent (UV-caused C > T mutations in melanomas, and tobacco-caused C > A mutations in lung carcinomas) which causes originally bulky lesions in the DNA that are repaired by NER. In contrast, no increment of the mutation rate in TFBS is observed in colon adenocarcinomas, where NER activity is not expected to play a major role in the mutational process, and only a modest increment is detected in other tumour types. Note that given the small number of whole-genome samples available and the lower mutational burden of breast, bladder and head and neck tumours compared to melanomas, lung carcinomas and colorectal tumours (Extended Data Table 1), the results for these tumour types should be taken with caution. Future analyses with larger cohorts of whole genomes, which would also allow a more accurate and specific separation of mutations by mutational processes should shed clearer light on this question.

Extended Data Table 1 | The whole genome sequencing data of different cancer types from TCGA and matching primary cell types from the Epigenome Roadmap

Cancer Type	Number of whole genome samples	Total number of mutations*	Mutations per sample (median value)*	Matching primary cell type from Epigenome roadmap (EID) or ENCODE cell lines†
Skin Cutaneous Melanoma (SKCM)	38	3,336,384	46,600	Foreskin Melanocyte Primary Cells (E059)
Lung adenocarcinoma (LUAD)	46	1,030,242	14,166	Fetal lung (E088)
Lung Squamous cell carcinoma (LUSC)	45	1,404,152	32,357	Fetal lung (E088)
Colorectal (CRC)	42	3,556,383	11,594	Fetal Intestine Large (E084)
CRC - *Hypermutated*	8/42	2,909,900	404,986	Fetal Intestine Large (E084)
CRC – *Hypomutated*	34/42	646,483	9,792	Fetal Intestine Large (E084)
Breast cancer (BRCA)	96	510,191	4,098	Breast variant Human Mammary Epithelial Cells - vHMEC (E028)
Bladder cancer (BLCA)	21	354,274	12,252	ENCODE cell lines
Head and neck squamous cell carcinoma (HNSC)	27	252,057	4,331	ENCODE cell lines
Normal skin cell	1	71,120	71,120	Foreskin Melanocyte Primary Cells (E059)

* The somatic mutations were previously called by Fredriksson *et al*., 2014[19], and the numbers presented here correspond to the total number of mutations after filtering as suggested by authors (see Materials section for more details). In the case of colorectal cancer (CRC), the samples were divided into two groups: Hypermutated and Hypomutated, based on the presence of missense mutations in the DNA polymerase epsilon gene or not. In the case of normal skin cell, obtained from Martincorena *et al*., 2015[20], the number presented here corresponds to the total number of single nucleotide substitutions.

† For each primary cell type, the respective DHS identified by Hotspot algorithm (FDR 1%) were download from http://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/narrowPeak/. The DHS from ENCODE cell lines was obtained from http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeRegDnaseClustered/wgEncodeRegDnaseClusteredV3.bed.gz.

CORRIGENDUM

doi:10.1038/nature16177

Corrigendum: Hypoxia fate mapping identifies cycling cardiomyocytes in the adult heart

Wataru Kimura, Feng Xiao, Diana C. Canseco, Shalini Muralidhar, SuWannee Thet, Helen M. Zhang, Yezan Abderrahman, Rui Chen, Joseph A. Garcia, John M. Shelton, James A. Richardson, Abdelrahman M. Ashour, Aroumougame Asaithamby, Hanquan Liang, Chao Xing, Zhigang Lu, Cheng Cheng Zhang & Hesham A. Sadek

Nature **523,** 226–230 (2015); doi:10.1038/nature14582

In this Letter we omitted to include the accession number for our RNA-seq data. The data are deposited in the Sequence Read Archive database under the accession number SRP060713. Also, in our description of the basic characteristics of cycling cardiomyocytes using the CAG line (Fig. 2b), we found that their cell size tended to be smaller, although this was not statistically significant ($P = 0.08749$, two-tailed unpaired *t*-test). The *P* value is reported correctly in Fig. 2b, but the original text incorrectly stated that cell size was significantly smaller. Author R.C. could not be contacted for approval of the publication of this Corrigendum. These errors have been corrected in the online versions of the paper.

CORRIGENDUM

doi:10.1038/nature16178

Corrigendum: Mapping tree density at a global scale

T. W. Crowther, H. B. Glick, K. R. Covey, C. Bettigole, D. S. Maynard, S. M. Thomas, J. R. Smith, G. Hintler, M. C. Duguid, G. Amatulli, M.-N. Tuanmu, W. Jetz, C. Salas, C. Stam, D. Piotto, R. Tavani, S. Green, G. Bruce, S. J. Williams, S. K. Wiser, M. O. Huber, G. M. Hengeveld, G.-J. Nabuurs, E. Tikhonova, P. Borchardt, C.-F. Li, L. W. Powrie, M. Fischer, A. Hemp, J. Homeier, P. Cho, A. C. Vibrans, P. M. Umunay, S. L. Piao, C. W. Rowe, M. S. Ashton, P. R. Crane & M. A. Bradford

Nature **525,** 201–205 (2015); doi:10.1038/nature14967

In the first boldface paragraph of this Article, the global number of trees should be approximately '1.30 trillion' (rather than '1.39 trillion') for tropical and subtropical forests and '0.66 trillion' (rather than '0.61 trillion') for temperate regions. These errors have been corrected in the online versions of the paper. In addition, the global tree density map can be found at http://elischolar.library.yale.edu/yale_fes_data/1/.

RETRACTION

doi:10.1038/nature16523

Retraction: The structure of complement C3b provides insights into complement activation and regulation

A. Abdul Ajees, K. Gunasekaran, John E. Volanakis, Sthanam. V. L. Narayana, Girish J. Kotwal & H. M. Krishna Murthy

Nature **444,** 221–225 (2006); doi:10.1038/nature05258

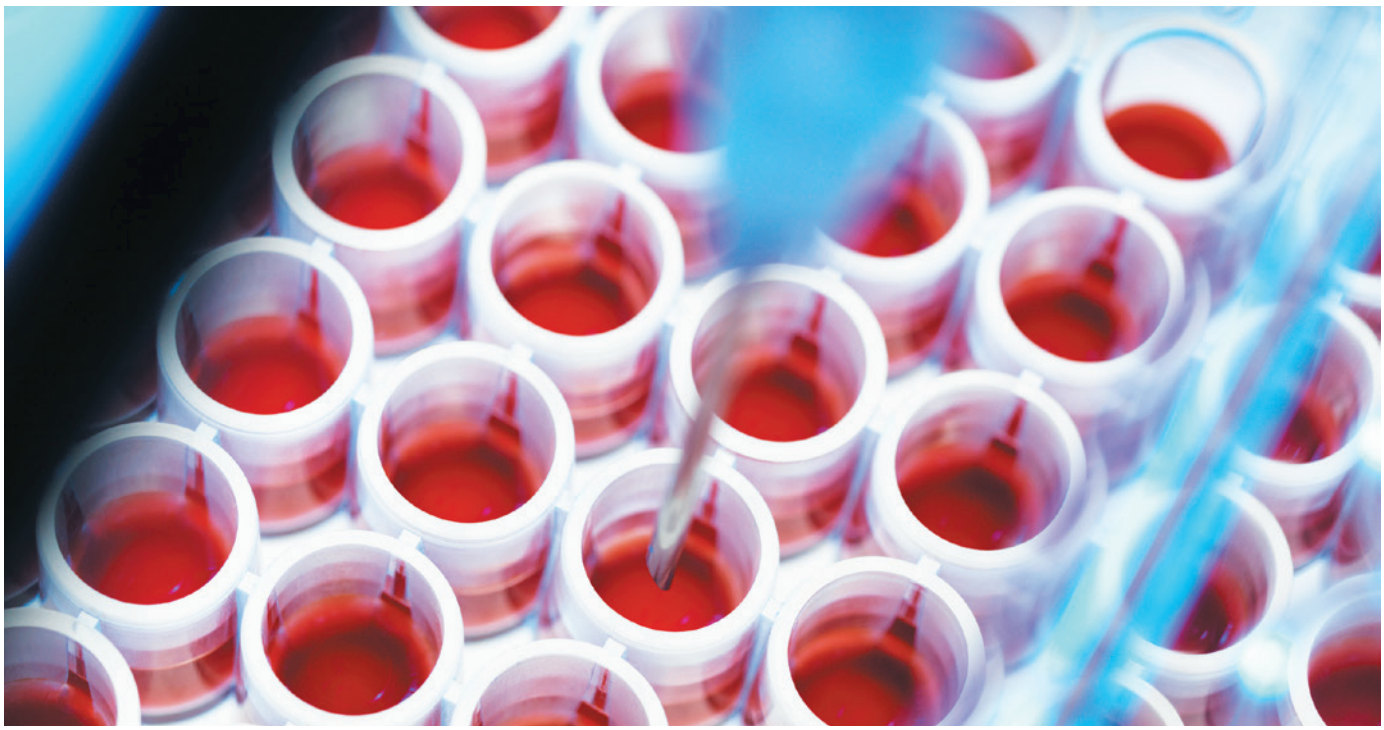
This Letter is retracted by *Nature*. This follows an investigation by the University of Alabama at Birmingham, Alabama, USA, of structures deposited into the Protein Data Bank under accession 2HR0 by H. M. Krishna Murthy. Co-authors who agree with the Retraction: A. Abdul Ajees, John E. Volanakis, Sthanam V. Narayana. Co-authors who do not support the Retraction: Girish J. Kotwal and H. M. Krishna Murthy. K. Gunasekaran has not responded. The report from the University of Alabama at Birmingham investigation is available at: http://www.uab.edu/reporterarchive/71570-uab-statement-on-protein-data-bank-issues.

TECHNOLOGY FEATURE

THE TUMOUR TRAIL LEFT IN BLOOD

Liquid biopsies can detect cancer signs from a blood sample, without the need for invasive procedures. But further work is needed before they can become reliable diagnostic tools.

CULTURA RM/ALAMY STOCK PHOTO



Tumour DNA extracted from blood samples could be used to profile cancers, avoiding the need for surgical biopsies.

BY KELLY RAE CHI

A lung biopsy is an invasive and uncomfortable procedure — especially for an 80-year-old grandmother. But by profiling his elderly patient's tumours in this way, lung oncologist Geoffrey Oxnard could target them with a matched drug. After treatment, his patient's tumours seemed to disappear.

Then, some time later, the 80-year-old returned to Oxnard's clinic riddled with pain. Tests showed that the cancer had returned, and hunting down a genetic cause of this resistance would require another invasive lung biopsy.

But Oxnard, who is at the Dana-Farber Cancer Institute in Boston, Massachusetts, offered the woman an alternative: "Let's just check your blood." He performed what's known as a liquid biopsy, using nothing more than a blood sample. Within a day, he spotted minuscule amounts of tumour DNA that revealed a mutation that causes resistance to treatment. Luckily, a drug that targets the mutation was being tested in clinical trials. With the genetic profile in hand, Oxnard managed to enrol his patient into the study, and her tumours went into remission again.

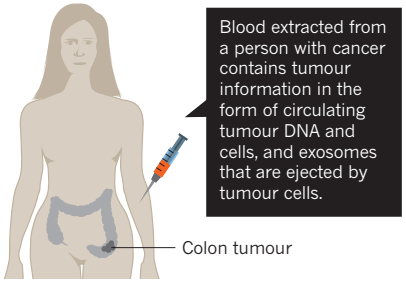
The discovery that parts of tumour cells, or even whole cells, break away from the original tumour and enter the bloodstream led to the idea of liquid biopsies. With this approach, cancers can be genetically characterized by analysing tumour DNA taken from a blood sample, thus bypassing the need to extract solid tumour tissue. Now, the rise of rapid genome-sequencing techniques has made it practical to translate this concept to the clinic. Three main approaches are being pursued: analysing circulating tumour DNA[1], examining whole tumour cells in the bloodstream[2] and capturing small vesicles called exosomes that are ejected by tumours[3] (see 'Scalpel-free biopsies'). And scientists have found that blood platelets might be able to offer up cancer clues, too (see 'Platelets ingest tumour data').

The allure of liquid biopsies is that they are quick, convenient and minimally painful, and they allow clinicians to closely monitor how tumours respond to therapies and to forecast cancer recurrences. In the long term, clinicians might even be able to use liquid biopsies to catch tumours at the earliest stages, before a person shows any symptoms. The genomic information in DNA circulating in the bloodstream could provide a snapshot of cancer genes in the body and may even point to where the cancer originated.

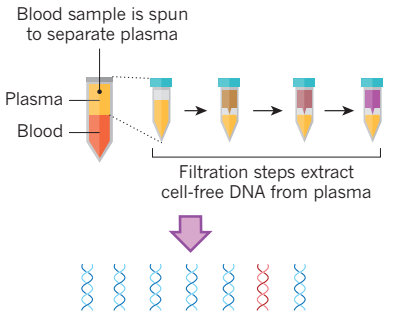
Investors are excited, and funds are pouring into start-ups focused on liquid biopsies. Sequencing firm Illumina of San Diego, ▶

SCALPEL-FREE BIOPSIES

Three different non-invasive techniques allow scientists to monitor tumours by performing 'liquid biopsies' on vials of blood.



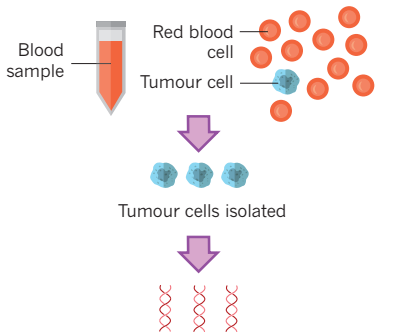
Circulating tumour DNA



DNA fragments from malignant cells (red) are separated from normal DNA (blue) and analysed by next-generation sequencing or digital polymerization chain reaction (dPCR).

Circulating tumour cells

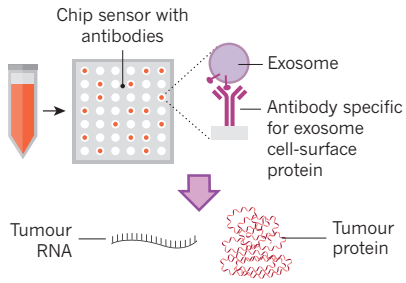
Circulating tumour cells are isolated from blood by cell-separation systems.



Cells are broken up to obtain tumour DNA that can be analysed by whole-genome sequencing.

Exosomes

Tumour exosomes are extracted from blood samples using different assays.



The material inside the captured exosomes — RNA and/or proteins — is then analysed.

California, for example, launched a spin-off company in January called Grail that will develop a plasma-based genetic screen for the early detection of multiple cancers.

But extensive testing is required before liquid biopsies can supersede surgical biopsies in the clinic. And there are still concerns from regulators about the sensitivity and accuracy of these procedures — for example, Oxnard's patient was required to have another, conventional biopsy to confirm the results of the liquid biopsy before she was allowed to enrol in the clinical trial.

Nevertheless, researchers say that it is no longer a question of whether liquid biopsies will one day replace surgical biopsies, but when and in what form. First, however, costs need to go down and sensitivity needs to rise.

FREE-FLOATING DNA

Bits of DNA are constantly flooding into the bloodstream. This genetic flotsam is present even in healthy people and could come from anywhere in the body. Dennis Lo, a chemical pathologist at the Chinese University of Hong Kong, realized that if the placenta of pregnant mothers released fetal DNA, then tumours may also shed DNA. Lo pioneered non-invasive prenatal screening for identifying chromosomal abnormalities in unborn babies, a test now in widespread use. But these screens have also yielded unexpected information — a few mothers-to-be found out that they had cancer.

The advent of more-accurate next-generation sequencing has enabled liquid biopsies on free-floating tumour DNA, because it can distinguish such sequences from normal DNA. In addition, digital polymerase chain reaction (dPCR) allows researchers to detect or quantify specific stretches of tumour DNA even at levels as low as 0.1% of total DNA in the blood.

"These are very sensitive tests," says dermatologist David Polsky of the New York University Langone Medical Center, who has shown that droplet-based dPCR can be used to monitor metastatic melanoma after treatment by tracking circulating DNA[4]. "It's because they're so impressive in the lab that clinicians are excited about them," he says.

Digital PCR, especially the droplet-based versions, has become so easy and cheap that many liquid-biopsy assays have the potential to become widely adopted. But the drawback of dPCR is that clinicians must know what aberrant DNA sequence they are looking for. As a result, sequencing could be preferable for some clinical indications because it allows clinicians to search for mutations without any pre-conceived notion of what genetic changes might be driving the cancer.

At the moment, liquid biopsies are mostly confined to basic- and clinical-research settings, although some blood tests are creeping onto the commercial market. Pathway Genomics, a medical-diagnostic company in San Diego, California, came under fire from the US Food and Drug Administration (FDA) in September 2015 for marketing its blood-based test directly to consumers. The regulator said that the CancerIntercept Detect screening test, which costs US$699 and is aimed at people who are at high-risk of developing cancer but are otherwise healthy, was not approved for direct marketing to consumers and had not been adequately clinically validated. The company countered that it had physician involvement and did not follow a direct-to-consumer model.

Lo is now finishing a 20,000-person screen for nasopharyngeal cancer, a rare type of head and neck cancer that is nevertheless common among men in southern China. In a 2013 study, Lo's team screened 1,300 healthy individuals and found 3 with early-stage cancer (stage 1). Lo will soon report the full results of his study. "It's a big deal that we are able to get them at stage 1," Lo says, because 95% who are treated at that stage survive. Those three people were promptly treated and are still healthy, he adds.

It helps that nasopharyngeal cancer is particularly easy to spot. The cancer is caused by Epstein–Barr virus, which leaves behind a specific genomic footprint that a simple, dPCR-based assay costing roughly $25 per test can pick up. But most other cancers have more undefined patterns of mutation. And although scientists can now access a growing catalogue of tumour signatures thanks to large-scale cancer-genome sequencing projects, there are many more to find, says Lo.

To track the source of the DNA fragments, scientists are starting to take size into account. Of the DNA bits floating in the bloodstream, those derived from normal cells are roughly 100–200 base pairs long, and are still wound around proteins called histones. Histones package DNA into the nuclei of different cells in different ways, so the length of the DNA may indicate the organ from which it was derived. Tumour DNA (like fetal DNA) is shorter than normal blood DNA fragments across multiple types of cancer, and large concentrations of very short fragments have been linked to metastasis (the migration of cancer through the body). The presence of tissue-specific transcription factors and other markers can also reveal clues about the tumour's provenance[5]. With all of this genetic information, clinicians might be able to make an informed guess as to where to look for a tumour, Lo says.

SPOTTING RELAPSE

Tracking a patient's circulating DNA also opens up the game-changing possibility of detecting metastasis at the very early stages, something that would otherwise require repeated invasive procedures. In the past few years, scientists have been studying circulating DNA for signs of cancer recurrence in women who were treated and presumed cancer-free. This is an important step towards early detection of metastatic relapse, says Muneesh Tewari, an oncologist and researcher at the University of Michigan Health System in Ann Arbor.

ADAPTED FROM C. B. MEADOR & C. M. LOVLY *NATURE MED.* **21**, 663–665 (2015); M. R. SPEICHER & K. PANTEL *NATURE BIOTECHNOL.* **32**, 441–443 (2014)

Platelets ingest tumour data

Blood platelets are better known for doling out clotting factors after a scrape than for their eating prowess. So when he first set up his laboratory at the VU University Medical Center in Amsterdam, Tom Würdinger was astonished to see platelets swallowing vesicles loaded with tumour RNA.

If platelets, which are easily isolated and counted in everyday blood tests, can take in transcripts from tumours, they might well provide a diagnostic treasure trove, Würdinger reasoned. He and Jonas Nilsson, then a postdoctoral researcher in the lab, formed the company thromboDx and analysed the RNA in platelets from people with cancer; they found that the RNA profiles looked markedly different in cancer, even in the 39 people who had early-stage cancers.

Würdinger's team has since tested the technology on nearly 1,000 people with 10 different types of cancer. More lab studies need to be done, along with trials assessing clinical utility. Still, Würdinger says that in liquid-biopsy development, "we cannot ignore platelets any more, because the results are so powerful". He adds that platelet tests may work well in combination with tests on other biomarkers such as circulating DNA. Indeed, sequencing company Illumina of San Diego, California, acquired thromboDx in February as part of its move into liquid biopsies. K.R.C.

In a 2015 study[6], a blood test in 20 people with breast cancer revealed signs of metastasis 3 years before it could be diagnosed with standard clinical tools. The biomarker was chromosomal rearrangements in circulating DNA. These rearrangements seem to occur early in cancer development, and scientists can use them to design primers for running dPCR tests.

"We thought that [circulating tumour] DNA should be a good biomarker for recurrence of breast cancer, but it had not really been shown before. When we saw that, it was very exciting," says Lao Saal of Lund University in Sweden, who led the study. "To think that you could pick it up three years beforehand, one could speculate that perhaps that could make a clinical difference."

"These are very sensitive tests. It's because they're so impressive in the lab that clinicians are excited about them."

Another study[7] of 55 people with breast cancer identified relapse about 8 months before symptoms appeared, also using dPCR to detect mutations. "Twenty per cent of women with breast cancer will go on to die of their cancer," says Nicholas Turner of the Institute of Cancer Research in London, whose team published the results. "We need to get an awful lot better at identifying the 20% who aren't cured by their current treatment."

Both Saal and Turner compare DNA from patients' primary tumours with circulating DNA collected after treatment. Saal has formed a company, SAGA Diagnostics in Lund, Sweden, that expects to partner with cancer specialists to validate the assay in the clinic. Saal is also working to improve the sensitivity of dPCR for identifying other point mutations to diagnose and monitor disease. Turner's group is starting a clinical trial to assess whether its assay can predict a person's response to a new class of immunotherapeutic drugs, the 'checkpoint inhibitors', which kick start the body's immune system.

Although liquid biopsies are beginning to gain a foothold in some clinics, it's still unclear what impact the search for pre-defined genetic alterations will have on patients' lives, Tewari says. Tumour resistance can emerge from mutations that are not covered by today's tests, so next-generation sequencing approaches will probably be needed to monitor all possible cancer mutations, Turner says.

BREAKAWAY CELLS

Whereas free-floating tumour DNA was discovered fairly recently, scientists first reported finding tumour cells in the blood in 1869. More than 40 different devices for isolating tumour cells are described in the literature (although only one, CellSearch by Veridex in Raritan, New Jersey, is approved by the FDA). For a long time, scientists pursued the notion that the number of tumour cells in blood might be used to assess the aggressiveness of a patient's cancer.

However, results at the bedside have been disappointing. In 2014, for example, the National Cancer Institute's Southwest Oncology Group published a study[8] showing that it was possible, by capturing and counting tumour cells in the blood of people with metastatic breast cancer, to identify a subset of those individuals with a more aggressive cancer. But although these women received another round of chemotherapy, this did not improve their outcomes. From this, people concluded that counting tumour cells had no practical value as a predictor, says Stefanie Jeffrey, a clinical oncologist at Stanford University in California. But, she adds, this conclusion was too swift: rather, it was the extra treatment that did not work. Nevertheless, the American Society of Clinical Oncology does not recommend counting cells to help guide treatment.

Rather than just counting tumour cells, a better option is to sequence the DNA or RNA from them to provide more insight into the gene alterations driving the patient's cancer. And because isolation techniques are improving, greater numbers of viable cells can be harvested from a patient — enough to allow researchers to culture the cells or implant them into mice to study their functional attributes. Any insights they gain can then be used to guide the patient's treatment. Klaus Pantel, director of the Institute of Tumor Biology at the University Medical Center Hamburg-Eppendorf in Germany, foresees the potential of this technique for predicting a tumour's response to targeted treatment — for example, researchers might spot PD-L1 on the cell surface, a protein that helps tumours to avoid the immune system and that can be targeted with immunotherapeutic drugs.

EXOSOME EXAM

A third liquid-biopsy approach targets exosomes. These are tiny vesicles that are shed by all living cells, including tumours, and just like their parent cells, they contain DNA, RNA and proteins. Much about them remains unknown, but a cancer test based on exosomes has already been commercialized. In January, a blood test developed by Exosome Diagnostics of Cambridge, Massachusetts, for detecting lung cancer by analysing tumour-exosome RNA was certified for laboratory use under the Clinical Laboratory Improvement Amendments (CLIA) quality programme in the United States. Exosomes can be harvested from a patient's blood, and potentially from other bodily fluids such as urine, which are even more convenient and easy to access than blood. However, the isolation of tumour-derived exosomes, which are variable in size, and their separation from normal exosomes, remains particularly challenging when compared with isolating free-floating DNA or tumour cells, so it will take some time before the technique is mature enough to detect other forms of cancer.

Each type of measurement gives a different window into the biology and course of disease, Tewari says. Much refinement of the techniques remains to be done, but clinicians are still excited about what liquid biopsies can do today. "They have a powerful role in helping patients get to the right treatment," says Oxnard. ■

Kelly Rae Chi *is a freelance science writer based in Cary, North Carolina.*

1. Bettegowda, C. *et al. Sci. Transl. Med.* **6,** 224ra24 (2014).
2. Lohr, J. G. *et al. Nature Biotechnol.* **32,** 479–484 (2014).
3. Im, H. *et al. Nature Biotechnol.* **32,** 490–495 (2014).
4. Chang, G. A. *et al. Mol. Oncol.* **10,** 157–165 (2016).
5. Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. *Cell* **164,** 57–68 (2016).
6. Olsson, E. *et al. EMBO Mol. Med.* **7,** 1034–1047 (2015).
7. Garcia-Murillas, I. *et al. Sci. Transl. Med.* **7,** 302ra133 (2015).
8. Smerage, J. B. *et al. J. Clin. Oncol.* **32,** 3483–3489 (2014).

TRADE TALK Mathematics whiz seeks real-world problems **p.274**

JUGGLING Balancing lab, baby and disability **go.nature.com/fhpxjr**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

ADAPTED FROM ABSCENT84

BIOMEDICAL RESEARCH

Privacy rules

Researchers must unpick a tangle of regulations to work with personal health data.

BY ALAINA LEVINE

Epidemiologist Thomas Sellers thought that he had discovered a treasure trove of data that could help to shed light on the heredity of breast cancer. When he took a post at the University of Minnesota in Minneapolis in 1989, he had heard about archived records of a multi-generation family study of the disease that were mouldering in the basement of the botany department. The study had been completed in 1952, and no one had kept track of the families that had been involved.

Excited by the prospect of data that spanned decades and generations, Sellers sifted through index cards that listed the names of people with breast cancer, and their relatives. He started to track down descendants of the patients to ask about their health and medical histories. "We ended up with four- or five-generation pedigrees," he says. "It was a really powerful resource."

But about seven years into his work, Sellers hit a major snag: in 1996, the US government passed the Health Insurance Portability and Accountability Act (HIPAA), a law that, among other things, established strict protections for the health information of individuals. His efforts to contact the relatives had to cease. "We were revealing information about people's cancer history to others who might not be allowed to know," says Sellers, now director and an executive vice-president of the Moffitt Cancer Center in Tampa, Florida. "It is a study that could not be done today."

Privacy laws have complicated research that involves people in many fields. Early-career investigators must navigate an ever-changing maze of regulations, but they do not have to face the challenge alone. Institutional review boards and compliance offices of universities and research centres can provide guidance on each step, from obtaining patient consent to handling and storing human tissue and data. Working closely with colleagues who are familiar with the issues — both within and beyond their institution — can also help researchers to get the data that they need without falling foul of the law.

BEWILDERING PATCHWORK

An important first step in many areas of biomedical research is for scientists to become familiar with the privacy laws that affect their work. In the United States, human-tissue research is governed mainly by two wide-ranging laws: the HIPAA and the Federal Policy for the Protection of Human Subjects, which is also known as the Common Rule. These laws dictate how researchers can obtain and use tissue and how they may store and protect the personal information that they collect.

Regulations vary widely between US states, and some state laws are tighter than federal laws; California, for example, has set a higher standard for medical privacy. And most institutions will also have their own policies and procedures, which can create a bewildering patchwork of requirements, especially for researchers who are part of multi-institution collaborations. "Cancer research in the United States is a fragmented effort," says Melissa Markey, a lawyer with Hall, Render, Killian, Heath and Lyman in Troy, Michigan, who specializes in technology, privacy and human-subject research. "This is the reason that researchers run screaming from explanations of how these laws fit together, because it is very confusing. It's like Alice in Wonderland."

Rules and responsibilities also vary from nation to nation. In the United Kingdom, the Data Protection Act controls the use of personal information, and the Human Tissue ▶

TRADE TALK
Serial solver


IBM RESEARCH

Cristiano Malossi works at the IBM Laboratory in Zurich, Switzerland. He won an IBM Research Prize for his PhD thesis in 2013 and the ACM Gordon Bell Prize two years later. He has used his mathematical skills to model blood flow, design aircraft, simulate convection in Earth's mantle and improve energy efficiency in high-performance computing.

Did you always want to work in industry?
I did my bachelor's and master's degrees studying aerospace, and my PhD in applied mathematics. Since the beginning of my studies, I was oriented towards a job in a company with high impact on technology. I wanted to see my work applied towards real products and services. The problems you solve in academia are generally more fundamental and long term; in a business environment, you are exposed to many different ones every day, and you are asked to get them solved almost as soon as you get them. This drives you to learn a lot of different things, very fast.

How did you find the perfect position?
At the end of my PhD, I didn't have time to send out CVs, and my professor said, 'if you want to stay as a postdoc, we welcome you', and so I stayed. I invested six months in sending out the best possible CVs to find the place I wanted to work. When you compete for the most prestigious positions, you can't send the application after half a day of working on it.

How did applying for prizes aid your search?
You show that you have high targets and that you are a great teamworker. Even if you don't win, the fact that you participated shows that you are motivated and that you want to do more than what is expected from you, which is what many companies are looking for. And I must admit, I was lucky; winning the IBM research prize for my PhD thesis on algorithms that allow patient-specific simulations of blood flow in the arteries was a strong way to get in touch with the people here. ■

INTERVIEW BY MONYA BAKER

This interview has been edited for length and clarity. See go.nature.com/li3gbs for more.

Act (and its counterpart in Scotland) regulates the use of human organs and tissues. The National Health Service (NHS) helps to direct how personal medical information can be shared. Senior members of staff in the NHS act as 'Caldicott guardians' who work to ensure that those data stay secure. "That seems like a lot of regulations, and it is," says Stefan Symeonides, a clinical oncologist at the University of Edinburgh. "My advice is to not be daunted. It's a lot of process, but the underlying principle is to enable research and maximize use of data in a safe way."

In Europe, harmonized laws facilitate the flow of tissues and data between EU member countries and beyond.

KNOTTY PROBLEMS

It can be a challenge to navigate the acquisition of health data. Large institutions and academic health-science centres in the United States and the United Kingdom typically employ or retain individuals who have expertise in privacy law and can offer comprehensive support to researchers. Madhu Purewal, a senior legal officer at the University of Texas MD Anderson Cancer Center in Houston, earlier this year helped an investigator to procure patient data from a handful of institutions that had different protocols. She guided the researcher in crafting individualized agreements for each. "As a faculty member, this is not your area of expertise," she says. "But I can help you figure out what is needed."

Carlos Caldas, an oncologist at the University of Cambridge, UK, says that he and his colleagues rely heavily on their institutions' clinical-research coordinators and data-security staff to steer them through the regulatory requirements. His advice? "Join places that have a critical mass of expertise." Caldas also says that large cancer-research facilities tend to have the infrastructure — tissue and tumour banks and encrypted databases — to accept and process samples without putting materials or data at risk.

Launching a research programme at an institution with no affiliation to a hospital can be a trickier matter. The lack of access to patients created cumbersome obstacles for biomedical engineer Michael Fenn, who works on cancer diagnostics. As a new member of faculty in 2013 at the Florida Institute of Technology in Melbourne, Fenn's research stalled when he tried to get patient samples and data from other research centres. The institute had no formal partnerships with cancer hospitals or research institutes, so he was uncertain about whom to contact at those organizations or how to comply with their privacy requirements. "I'm asking them, 'May I have some tissue from a particular type of patient?'" he says. "But the process was so convoluted and I wasn't even sure how to initiate it."

Fenn smoothed the way by establishing relationships with key researchers, surgeons and pathologists at the centres who helped him to navigate the process of accessing tissue and data. He now advises early-career researchers to establish informal alliances before they even think about getting their first samples. "You'll get the access you need, and the scientists and physicians there will help you to move beyond the bureaucracy," he says.

Even investigators at large cancer-research centres are likely to encounter bureaucratic knots, particularly when participating in large collaborations that span institutions. "In an era of team science, that can be really difficult," Sellers says. "There might be 30 institutional review boards involved for one study, each with their own agreements. It takes time, money and effort, and it's not helping to accelerate academic health-related research."

Individual researchers can be frustrated by restrictions. Katerina Politi, a pathologist at the Yale Cancer Center in New Haven, Connecticut, has obtained consent from patients that allows her to collect their tissue for immediate analysis. But patients must provide further consent if another sample is needed from them. "We can do this biopsy, but if they have another in the future, they have to reconsent," says Politi. "If you could streamline the consent of patients and acquisition of materials, you might not miss opportunities to learn more about diseases."

Joining forces with peers and colleagues is the best way to untangle the knotty problems of privacy.

Some institutions are trying to smooth the process. In 2014, oncologist Michael Caligiuri, who directs the Ohio State University Comprehensive Cancer Center in Columbus, in 2014 co-founded the Oncology Research Information Exchange Network, a federation of 11 cancer-research centres across the United States. Member organizations share an institutional review board and follow a uniform protocol for interacting with patients and requesting and collecting tissue and data. Caligiuri says that studies performed within the network can move more quickly and require less paperwork because members can share data and samples from patients.

Ultimately, joining forces with peers and colleagues is the best way to untangle the knotty problems of privacy, say seasoned researchers. "The rules are always changing," says Sellers. "One needs to be paying attention to the literature and what's coming out there from the government. Find people in your network who are dealing with the same challenges. We're always happy to share our recipes for obtaining the data and consent." ■

Alaina Levine *is a freelance science writer based in Tucson, Arizona.*

INCURSIONS

Report from the front line.

BY LAVIE TIDHAR

ILLUSTRATION BY JACEY

The wardrobe had sat in the old house by the seaside for generations. Who bought it, we didn't know. Its timber came from our far-flung colonies in the South Pacific, places that have long since gained their independence from us, at least by letter of the law. It had survived the bombings in the Second World War and subsequent ownerships, recessions, booms, property bubbles and changing governments.

All through this time the sea continued to beat against the shore, futilely, and the wardrobe just sat there, unchanged, made of old weathered teak and oak and pine. Often, when you opened it, you would feel a cold draught blowing, and hear the rustle of pines, but it was an old house and such things are not unheard of. A previous owner stored old fur coats there, but those had long since rotted away and were thrown on the rubbish heap. Nowadays, it just stored children's clothes bought off the rack.

The first incursion occurred at a quarter past one at night. There were no witnesses but for a young girl, clutching her teddy bear, who opened one eye in her bed and glanced into the darkness and then looked away. The faun (it was an older specimen than the ones later caught) tiptoed through the room and out of the door. When it was first spotted it was some two hours later, on the road leading to London. Witnesses later said that the faun looked disorientated, confused. The glare of passing lights frightened it.

The element of incursion (sub-species Pan) was eliminated three hours and forty-seven minutes after entry. In his testimony, the lorry driver who hit the intrusion said, "He just stood there. I had no time to brake. He just stood there as though he'd never seen a truck before."

We can confirm that the specimen was male. It was vaguely human-shaped, with cloven hooves, reddish skin, curly hair and horns. An extensive autopsy carried out later concluded that it died of severe injuries caused on impact with the moving vehicle.

There was a delay as the information filtered slowly through the ranks of law-enforcement and government agencies until reaching my department. This was unfortunate in light of later events. By the time we had set up a perimeter around the house, an unknown number of entities had escaped into the English countryside. From later, partial reports, we know these included: centaurs (category: Hybrid), dwarfs, giants, nymphs, two unicorns, a tribe of talking mice and assorted other species, and at least one river god, who hid itself in the Severn. It took an eight-day battle with the army before the entity was successfully eliminated.

Our first real break was when a farmer in East Anglia managed to catch one of the badgers. Although we had placed the entity under interrogation, it initially refused to talk. When it did, it — like the first incursion — seemed confused. It insisted the entities were refugees, fleeing some sort of oppressive regime in their land. Our interest was aroused at the mention of a possible WMD, a device capable of causing permanent winter. Such a thing is beyond our current tactical capabilities. We established a working group to look into the military implications as we sought simultaneously to contain the incursion and round up the intruders into secure facilities, or camps, an operation that was, on the whole, successful.

There were, however, complications.

A troupe of travelling fauns was spotted around Piccadilly Circus, begging for scraps from passers-by. They proved to be surprisingly fleet-footed, and avoided capture for several months.

An assembly of nymphs took root in Epping Forest. They assaulted joggers and ramblers and violently resisted any incursion by our forces. Several giants took residence in Surbiton, on the hill, and talking mice bred like the plague in some parts of Manchester and the northwest.

Local police forces, ill-equipped for this sort of thing, found it hard to cope.

The family who inhabited ground zero were evacuated from the house on the first day of the incursion. We had set up a perimeter to prevent any further unauthorized access, but refugees continued to slip through. Worse — we began to suspect that their tales of a cruel monarch and her despotic rule were not altogether untrue. Hostile elements successfully planned and executed an attack — using unknown means — which turned half of Rickmansworth into a block of ice (though it was hard to tell at first). A series of ritual killings, targeting lions, shocked the nation, leaving behind it dead and mutilated animals in every zoo in the country.

We reluctantly began to discuss countermeasures.

The wardrobe has been sitting in the house for generations. Who bought it we didn't know. Its interior smelled of pine needles and ice, of lichen and washing powder. We continue to send men into the forbidden zone, although few come back, and those who do are changed by age and years, despite the fact that little time passes beyond the perimeter fence.

As for the refugees, they continue to come. We attempt to contain them, and the camps, daily, continue to grow. The wardrobe has been sitting in the house for generations and could last, at a conservative estimate, for generations more.

Who bought it, we still don't know. ■

Lavie Tidhar *is the author of the Jerwood Fiction Uncovered Prize winning* A Man Lies Dreaming*, the World Fantasy Award winning* Osama *and the critically acclaimed* The Violent Century. *His other works include the* Bookman Histories *trilogy, several novellas, two collections and a forthcoming comics miniseries,* Adler. *He currently lives in London.*

NATURE.COM
Follow Futures:
@NatureFutures
go.nature.com/mtoodm